eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

TESIS DOCTORAL

Utilización de la fase armónica en la detección de voz sintética

Jon Sanchez de la Fuente

Directores:
Dra. Inmaculada Hernáez Rioja
Dr. Ibon Saratxaga Couceiro

Aholab Signal Processing Laboratory
Departamento de Ingeniería de Comunicaciones
Universidad del País Vasco/Euskal Herriko Unibertsitatea

Bilbao, 2016

A las tres 'jefas'

Agradecimientos

Cuando se ha estado tanto tiempo desarrollando un trabajo, se colabora con tanta gente que a la hora de escribir estos agradecimientos asusta la posibilidad de olvidarse de alguien. Confío en que en esta ocasión no suceda.

Quiero empezar agradeciendo a mis directores el acompañamiento que me han brindado durante todos estos años.

Inma es una líder con todo el alcance de la palabra. Su energía y su capacidad de trabajo arrastran a todo el grupo Aholab, e indudablemente yo no soy una excepción. Si alguien ha hecho posible que esta tesis llegue a existir, sin duda es ella.

Ibon también ha sido guía imprescindible durante los últimos años, y el tiempo en que he compartido despacho con él ha sido en el que más rápidamente ha avanzado esta investigación. Esos debates frente a una pizarra de cristal a última hora del día pueden ser sorprendentemente fructíferos.

También el resto del grupo Aholab ha aportado su granito de arena para que este proceso haya podido avanzar.

Eva es una compañera que cualquiera querría tener. Una dosis de optimismo y vitalidad que levanta el ánimo incluso en los momentos de mayor desesperación. Con un gusto para la estética que puede hacer que cualquier póster parezca diseñado por un profesional, y un conocimiento de las tecnologías del habla que puede dar una visión fresca a cualquier línea atascada.

Dani ha sido de gran ayuda aportando materiales para las bases de datos, descripciones de los sistemas y, sobre todo, una visión práctica de la “política de investigación” que me ha ayudado a entender muchos mecanismos.

Igor, David, Agustín y Luis han mostrado su apoyo incondicional en cada proyecto y cada comida que hemos compartido. Además de llevar con alegría el arreglo de

problemas con el servidor y las contiendas por el reparto de la capacidad de cálculo y de disco.

También el grupo del café ha ayudado, dando su apoyo y aportando ideas y visiones. Sé que todos ellos van a alegrase de ver esta tesis terminada. En general, para todo el Departamento de Ingeniería de Comunicaciones será una alegría.

Iker hace años que salió de Aholab y cruzó el Ebro. Pero incluso desde la distancia ha estado siguiendo el avance de la investigación y dando consejos, que han resultado ser muy valiosos.

También Iñaki S. jugó un papel importante en la generación de parte del material. Espero que llegue a leer estas líneas de agradecimiento, y que consiga sus objetivos vitales.

Pero no menos importantes han sido los apoyos recibidos fuera de la Escuela. Creo que en este ámbito Itziar es quien mayor reconocimiento merece. Le ha tocado soportar las épocas más frustrantes de todo el proceso, a pesar de lo cual nunca ha perdido la fe en que esto llegaría a buen fin, ni ha aflojado la presión, que también es una forma de apoyo. Todo llega, y lo que falta, también llegará pronto. Seguro.

Mis padres y mis tías han hecho un esfuerzo considerable para entender que en la universidad, aun cuando no tenga clases, estoy para investigar. Solamente eso ya sería de agradecer, pero además nunca han dejado de interesarse, y nunca han perdido la ilusión.

Javi no ha dejado en ningún momento de seguir el proceso, intentando ayudar con el planteamiento, y elaborando planificaciones que una y otra vez modificábamos. Al final ha habido una que ha sido definitiva.

A Iñaki V. le ha tocado leer. Artículos en Lisboa. Resultados en Kyoto. Capítulos enteros en Bilbao. Siempre con interés, y siempre manteniendo la tensión. Pero no ha sido el único lector. Jose también ha leído algún capítulo, y no ha dejado de apoyarme aunque no siempre haya compartido lo que hago. Raquel ha ayudado en la introducción, y en algún otro aspecto de la vida. A Alfredo le ha caído alguna galerada. A Maribi alguna traducción. A Vero hacer propuestas sobre el diseño.

Hay mucha más gente que ha demostrado interés, y se alegrará sinceramente de ver el final de este trabajo. La SAP. La cuadrilla del Isusi. La Asociación. La Plataforma. El Orfeón. La familia política.

A todos, los mencionados aquí y los que (espero que no) me haya podido olvidar, muchas gracias por el seguimiento y el apoyo.

Resumen

Hoy en día, resulta crucial en muchas aplicaciones gestionar quien tiene acceso a determinados lugares o informaciones. Y desde hace tiempo, está cobrando fuerza una tendencia que invita a utilizar como elemento de acceso características biométricas en lugar de tarjetas, llaves, o claves. Las características biométricas tienen la ventaja que no pueden ser olvidadas ni robadas. Entre ellas, la voz es una de las más útiles, por ser claramente diferenciadora y poder obtenerse de forma poco intrusiva.

Los sistemas de verificación de locutor utilizan la voz como vector biométrico, y tienen que enfrentarse a la posibilidad de ser atacados mediante técnicas de spoofing. Hoy en día, las tecnologías de conversión de voces y de síntesis de voz adaptada a locutor han avanzado lo suficiente para que pueda plantearse que voces artificiales, sintetizadas o transformadas, sean capaces de engañar a un sistema biométrico de identificación de locutor.

En esta tesis se propone un módulo de detección de habla sintética (SSD) que puede utilizarse como complemento a un sistema de verificación de locutor, pero que es capaz de funcionar de manera independiente. Lo conforma un clasificador basado en GMM, dotado de dos modelos diferentes, uno de habla humana y otro de habla sintética. Cada señal candidata se compara con ambos modelos, y, si la diferencia de verosimilitudes supera un determinado umbral, se acepta como humana, rechazándose como sintética en caso contrario. El sistema desarrollado es independiente de locutor.

Los parámetros de la señal de voz que se utilizará para la generación de los modelos estarán basados en la fase armónica de la señal de voz, mediante la parametrización RPS.

Por otro lado, se propone una técnica para reducir la dificultad intrínseca del proceso de entrenamiento del sistema. Se busca evitar la necesidad de generar un sistema completo de adaptación de voces a las de los locutores del sistema o un juego completo de voces adaptadas a ellos para sintetizar mediante TTS. Para ello, y

sabiendo que la mayoría de los sistemas de adaptación o síntesis modernos hacen uso de vocoders, se propone transcodificar señales humanas mediante vocoders para obtener de esta forma sus versiones sintéticas, con las que, a su vez, se generarán los modelos de habla sintética del clasificador. Se demostrará que, detectando el vocoder intrínseco a las señales, el sistema es capaz de detectar su naturaleza sintética.

El rendimiento del sistema se pone a prueba en diferentes condiciones: primero, con las propias señales transcodificadas, utilizando modelos creados con distinto número de locutores, y con vocoders únicos o múltiples. Después, con ataques sintéticos creados mediante TTS. Por último, se plantean distintas estrategias para el entrenamiento de los modelos a utilizar en sistemas SSD.

Los resultados obtenidos ponen de manifiesto que con el sistema propuesto es posible detectar ataques de spoofing basados en voces sintéticas, utilizando modelos creados exclusivamente mediante parámetros RPS, y generados usando vocoders en lugar de ataques reales.

Abstract

Nowadays it is critical for some applications to handle with the access people have to some places or information. In the last years there is a growing tendency on using biometric characteristics instead of access-cards, keys or keywords. Biometric characteristics have one main advantage: they cannot be forgotten or stolen. Among all biometric vectors, voice is particularly appealing, as it can be clearly used to discriminate, and users feel largely comfortable about it.

Speaker Verification systems use voice as biometric vector. But an impostor could try to deceive the system by impersonating another enrolled user, by means of spoofing techniques. The development of voice conversion (VC) and text-to-speech (TTS) systems has acquire such a quality level that it is possible to create artificial voices, either converted or synthesized, than can fool a biometric speaker verification system.

In this thesis a Synthetic Speech Detection (SSD) system is proposed. It can be used to complement a speaker verification system, but works independently. It is based on a GMM classifier with two different models: a human speech model and a synthetic speech model. Every candidate signal is compared with both models, and, if the likelihood difference is greater than a given threshold, is accepted as human. Otherwise, it is rejected and taken for synthetic. The developed system is speaker independent.

The voice parametrization used to create the human and synthetic models are based on the harmonic phase of the voice, using the RPS parameterization.

Also, a technique is proposed to simplify the system training process. In order to generate the synthetic models, it is necessary to create specific VC or TTS system adapted to every speaker of the system. Knowing that most of these systems are based on vocoders, it is proposed that synthetic voice can be acquired by vocoding natural voices. That vocoded voices will be used to create the synthetic models. It will be demonstrated that the synthetic nature of and attacking voice can be pointed by detecting the vocoder used to create it.

The performance of the system is tested under different conditions. First with the vocoded signals themselves, with models created with different number of speakers, or using single or multiple vocoders. Then, a test is performed with TTS-based synthetic attacks. Finally, some different model training strategies are analyzed.

The presented results demonstrate how it is possible to detect synthetic voice based spoofing attacks, using RPS parameter based models trained by means of vocoded speech instead of realistic attacks.

Laburpena

Gaur egun, ezinbestekoa da aplikazio batzuetarako leku edo informazio batzuk erabiltzeko baimena nork daukan kudeatzea. Azken urteotan tendentzia bat indartzen ari da: txartelak, giltzak edo klabeak erabili beharrean ezaugarri biometrikoak erabiltzea aldarrikatzen duena. Ezaugarri biometrikoen abantailarik handiena ahazteko edo osteko ezintasunean datza. Eta ezaugarri biometrikoen artean erabilgarrienetariko bat ahotsa da: identifikazioa gauzatzeko beste informazio dauka, eta erabiltzailearentzako erosoa den eran lor daiteke.

Hizlariaren egiaztaketa sistemek ahotza erabiltzen dute bektore biometriko moduan, eta posible da ezaugarri hori antzeratzea sistema iruzurtzeko, spoofing izeneko teknikak erabiliz. Gaur egungo ahots bihurketa eta sintesi teknikek hizlariaren egiaztaketa egiten duen sistema bat iruzurtzeko beste kalitate lortu dute.

Tesi honetan Ahots Sintetikoaren Detektagailu modulu bat (Synthetic Speech Detection, SSD) proposatzen da. Hizlariaren egiaztaketa sistema baten osagarri moduan erabil daiteke, baina baita independenteki. GMM sailkatzaile baten oinarritzen da. Horrek bi eredu desberdin erabiliko ditu, giza-ahotsarena eta ahots artifizialarena. Sistemara sartzen den seinale bakoitzak bi ereduekin egiaztatzen da; bien egiantz-arrazoien arteko aldea emandako ataza baten gainetik badago, gizaki batek ahoskatua izan dela suposatuko da. Bestela, makina batek sortu duela. Garatutako sistema ez dago hizlarien menpe.

Ereduak sortzeko erabiliko diren ahozko seinalearen parametroak ahotsaren fase harmonikoan oinarrituak egongo dira. RPS parametrizazioa erabiliko da hain zuzen.

Sistemaren entrenamendu-prozedura errazteko teknika bat ere proposatzen da. Sistemako hizlari bakoitzeko ahotz bihurketa edo sintesi moldatuaren moduak osatzeko beharra ekiditea bilatzen da. Horretarako, gaur egungo bihurketa edo sintesi sistema gehienek bokoderrak erabiltzen dituztenez, giza-ahotsak transkodifikatuko dira bokoderren bidez, haien bertsio sintetikoak osatzeko. Horiekin, ondoren, eredu

sintetikoak sortuko dira. Seinaleen bokoderra detektatuz, sistemak bere jatorri sintetikoa detektatzeko gai dela egiaztatuko da.

Sistemaren etekina egoera desberdinetan aztertuko da: lehen, transkodifikatutako seinaleekin, ereduak hizlari kopuru desberdinekin sortuz edo bokoder bakarra edo ugari erabiliz. Ondoren, benetako eraso sintetikoak erabiliko dira sistemaren gaitasuna aztertzeko. Azkenik, ereduak osatzeko erabil daitezkeen estrategiak planteatuko dira.

Lortutako emaitzekin argi ikusten da proposatutako sistemarekin posible dela ahots sintetikoan oinarritutako spoofing erasoak detektatzea, bokoderrak erabiliz entrenatu diren RPS parametroez osatutako ereduekin.

Índice

Glosario

AS: Amplitude Scaling

ASV: Automatic Speaker Verification

DCF: Decision Cost Function

DCT: Discrete Cosine Transform

DNN: Deep Neural Networks

DTW: Dynamic Time Warping

EER: Equal Error Rate

FAR: False Acceptance Rate

FRR: False Reject Rate

FW: Frequency Warping

GMM: Gaussian Mixture Model

HMM: Hidden Markov Models

HNM: Harmonic plus Noise Model

HNR: Harmonic-to-Noise Ratio

HTK: HMM Toolkit

HTS: HMM Toolkit Speech synthesis

IAIF: Iterative Adaptive Inverse Filtering

IDCT: Inverse Discrete Cosine Transform

LFCC: Linear frequency cepstral coefficients

LPC: Linear Prediction Coefficients

LPCC: Linear Prediction Cepstral Coefficients

LSF: Line Spectral Frequencies

LSP: Line Spectral Pair

MBROLA: Multi Band Resynthesis Overlap and ADD

MFCC: Mel Frequency Cepstral Coefficients

MGD: Modified Group Delay

MLSA: Mel Log Spectrum Approximation

MOS: Mean Opinion Score

MRRPS: Mel Regularized Relative Phase Shift

MVF: Maximum Voiced Frequency

NIST: National Institute of Standards and Technology

OLA: Overlap-and-add

PAD: Playback Attack Detection

PLDA: Probabilistic Linear Discriminant Analysis

RPS: Relative Phase Shift

SID: Speaker Identification

SLM: Sinusoidal Likeness Measure

SSD: Synthetic Speech Detection

STFT: Short Time Fourier Transform

STRAIGHT: Speech Transformation and Representation by Adaptive Interpolation of weiGHTed spectrogram

SV: Speaker Verification

SVM: Support Vector Machine

TTS: Text to Speech Synthesis

UBM: Universal Background Model

VAD: Voice activity detector

VQ: Vector Quantization

WSJ: Wall Street Journal

Índice de Tablas

Índice de figuras

1. Introducción

En la sociedad de la información, establecer claramente la privacidad de los datos y quien puede acceder o no a un determinado servicio es crucial (Jain et al., 2006). Por tanto, en muchas ocasiones, a la hora de que personas y sistemas interactúen se hace necesario que una aplicación establezca de manera segura con qué persona se está tratando. Y aunque cualquier tipo de información es susceptible de ser restringida a un número limitado de usuarios, hay algunos ámbitos en los que el control de acceso se convierte en algo crítico, como puede ser el de las operaciones bancarias.

Tradicionalmente, los métodos de identificación o verificación de identidad se han basado en tres principios: tener, conocer, o ser. *Tener* basa la identificación en la propiedad, en ser portador de un objeto, como pueden ser tarjetas, llaves o similares. *Conocer* implica ser conocedor de una cierta información, como en el caso de las claves secretas. El problema de ambos estriba en que los objetos pueden ser perdidos o robados, y las claves pueden ser escuchadas u olvidadas, de forma que la investigación orientada a seguridad ha tendido en los últimos años a la tercera vía. *Ser* se basa en conseguir la identificación en base a rasgos biométricos (Fierrez et al., 2009) (Jain et al., 2006).

Los rasgos biométricos son características físicas inherentes a cada persona, de forma que es imposible que sean robados u olvidados. El iris de los ojos, la cara, la firma (o en general la escritura manual), las huellas dactilares, la geometría de la mano, la forma de teclear y la voz son algunos de los rasgos biométricos más comúnmente utilizados en los sistemas de acceso a información.

La voz, concretamente, ha demostrado ser uno de los sistemas de acceso biométrico más útiles, ya que puede obtenerse de manera poco intrusiva y muy natural para el usuario, tanto presencialmente como por teléfono, y ya está siendo utilizado por algunas empresas de telefonía y de banca para facilitar a los clientes el acceso a sus transacciones (Beranek, 2013). La tecnología ha llegado a un punto de competencia tal que los sistemas de identificación de locutor, es decir, específicamente basados en voz, pueden utilizarse en ciertos casos en aplicaciones forenses y jurídicas de identificación (Drygajlo, 2007).

A pesar de la dificultad intrínseca que presenta el robar rasgos biométricos, en algunos casos sí que pueden ser copiados o falsificados. En el caso concreto de la voz como vector biométrico, existen varios sistemas que permiten intentar burlar la seguridad. Algunos son muy simples (como pueden ser imitadores humanos tratando que su voz parezca la de otra persona diferente) pero otros son muy avanzados y, utilizan tecnologías del habla punteras. Por ejemplo, la tecnología de conversión de voz, que permite transformar la voz de un hablante, llamado hablante origen, de tal modo que los oyentes la perciban como si fuera la de otro hablante, llamado hablante objetivo (Erro, 2008). O la síntesis de habla con voz adaptada, que consiste en modificar mediante unas pocas muestras la voz de un sistema TTS (Text to Speech Synthesis) para que suene de manera similar a la voz del hablante a imitar. Ambas pueden resultar amenazas reales a la seguridad de un sistema de acceso biométrico. La síntesis de voces adaptadas

Como medida de protección de un sistema biométrico basado en voz ante ataques como los recién descritos, se pueden tomar dos medidas. La primera pasa por aleccionar al sistema de verificación de locutor para que cuando llegue una señal sintética atacante, la tome como de un locutor incorrecto y consecuentemente deniegue el acceso. La segunda consiste en implementar junto con el sistema de verificación de locutor, un módulo independiente de detección de voz sintética (Synthetic Speech Detection, SSD). Éste tiene como función detectar cuándo la voz que entra al sistema ha sido generada mediante una de las técnicas mencionadas anteriormente, independientemente del resultado que ofrezca el sistema de verificación.

En esta segunda medida de protección, el desarrollo de un módulo separado de SSD, centraremos el desarrollo de esta tesis.

1.1. Motivación

En este trabajo se toma como punto de partida los trabajos realizado en Aholab (Saratxaga, 2011) y (De Leon et al., 2011). En ellos se trata, por un lado, sobre la representación de la fase armónica de la voz mediante parámetros denominados Relative Phase Shift (RPS), y por otra, de las aplicaciones que se pueden dar a dichos

parámetros, incluyendo la detección de voz sintética (Synthetic Speech Detection, SSD).

La fase en la voz ha sido tradicionalmente desechada desde que (Ohm, 1843) estableció su Ley Acústica de la fase propugnando que el oído humano captaba la respuesta frecuencial de los sonidos en magnitud pero descartaba la fase. Siguiendo esta base, en muchas aplicaciones de las tecnologías del habla, aun hoy en día, no se realiza un gran esfuerzo en modelar correctamente la fase, cuando no se descarta directamente. Esto puede utilizarse como un elemento distintivo que permita distinguir voz natural de voz procesada en un sistema SSD.

El sistema SSD presentado en (Saratxaga, 2011) y (De Leon et al., 2011) implementa la detección de voz sintética dependiente de locutor. Para cada usuario, se crea un par de modelos (natural y artificial), de manera que la decisión sobre la naturaleza humana o sintética de una voz de entrada en concreto va ligada a la decisión sobre la identidad del locutor. En los experimentos realizados, se generan mediante HTS (HMM Toolkit Speech synthesis) (HTS Working Group, 2002) (Yamagishi et al., 2009) voces sintéticas adaptadas a cada locutor, que se utilizarán para realizar ataques al sistema. Los resultados obtenidos son de un 100% de acierto en la tarea de detección de voz sintética.

Sin embargo, el sistema descrito presenta dos problemas. El primero, que es un sistema totalmente dependiente de locutor, por lo que las decisiones sobre la naturaleza sintética o no de una señal de entrada solo serán válidas para los locutores presentes en el sistema. El segundo, que es necesario el desarrollo de tantos ataques adaptados como locutores se recojan en el verificador. Con ellos se entrenarán los modelos del sistema. Será necesario además, generar diferentes tipo de ataques para que los modelos puedan recoger todos ellos y se evite que la parte de detección de voz sintética esté vinculada al sistema de ataque utilizado. Todo ello implica que sea necesario generar señales de todos los locutores y con una gran diversidad de técnicas, complicando la viabilidad de su obtención.

En esta tesis se va a profundizar en el desarrollo de sistemas SSD. Manteniendo la información de fase como parámetro decisorio, y el modelo basado en parámetros RPS

que ha dado buenos resultados, se busca desarrollar un sistema más general, que sea capaz, no solo de detectar voz adaptada usando HTS, sino un abanico más amplio de posibles ataques, incluyendo otros sistemas de síntesis y la conversión de voces.

1.2. Objetivos de la tesis

El objetivo principal de esta tesis es **crear un sistema SSD universal**, capaz de detectar cualquier tipo de ataque, y segregado del verificador de locutor. Para ello se desarrollarán los siguientes objetivos parciales:

- Validación de la idoneidad de la parametrización RPS como base de un sistema de detección: se diseñará el sistema SSD en base a los parámetros RPS y se evaluará el rendimiento en la tarea de detección, comparándolo con un sistema de referencia basado en parámetros MFCC (Mel Frequency Cepstral Coefficients).
- Independencia de locutor: se pretende crear un sistema que sea independiente de locutor. Con esa premisa, se crearán modelos con distintos locutores, y con distinto número de locutores, y se analizará el rendimiento del sistema que los usa.
- Validación del uso de la técnica de transcodificación mediante vocoders para generar ataques: para sortear la necesidad de disponer de ataques de spoofing con que entrenar un sistema de detección, se propone el uso de vocoders. Con ellos se crearán versiones sintéticas transcodificadas de las voces humanas originales. El proceso de creación del sistema se simplifica al no es necesario disponer de ataques de spoofing. Dado que la mayoría de sistemas de síntesis y conversión actuales están basados en vocoders, se propone que detectar la presencia de un vocoder permitirá decidir que una voz es sintética.
- Independencia del vocoder: una vez establecido que se van a utilizar vocoders para el entrenamiento del sistema, se busca que la detección sea independiente del vocoder, y que se pueda detectar por igual en la entrada del sistema ataques creados utilizando cualquier vocoder.
- Evaluación del SSD frente ataques reales: Por último, para validar los métodos y modelos desarrollados, se enfrentará el sistema a ataques que puedan darse en

situaciones reales, como pueden ser voz sintetizada mediante TTS o voz convertida.

1.3. Organización de la tesis

En el capítulo 2 de la tesis se presenta el estado del arte, donde se describen los avances en las técnicas que se utilizarán: se comienza con la verificación e identificación de locutor, el ámbito de las tecnologías del habla en que será de aplicación el sistema SSD que desarrollaremos. Se dan detalles a continuación del spoofing o simulación, las distintas técnicas que se pueden utilizar para burlar un sistema de seguridad biométrico basado en identificación de locutor: grabaciones, imitaciones, conversión de voz y síntesis de voces adaptadas. Seguidamente se detallan las características de los sistemas SSD y los distintos parámetros de la voz en que pueden basarse.

En el capítulo 3 se presenta el sistema SSD propuesto, describiendo su esquema básico de funcionamiento, el modo de parametrizar y modelar las señales, el clasificador y el material de entrenamiento, incluyendo los vocoders y las bases de datos utilizadas para la creación de los modelos. Se describe también cómo dotar al sistema de la capacidad para ser independiente de locutor.

Una vez descrito el diseño del sistema, en los capítulos siguientes se probarán diferentes estrategias de diseño del entrenamiento del sistema para optimizar su rendimiento en distintas situaciones, enfrentándose a ataques de diferente naturaleza.

En el capítulo 4 se analizará la independencia del sistema para con el vocoder utilizado para entrenarlo, empezando con los propios vocoders utilizados para la creación de los modelos del sistema. Se establecerá la capacidad de los modelos creados con un vocoder para detectar ataques creados con otro vocoder diferente, para analizar su dependencia. Se ensayarán también estrategias para la creación de modelos que puedan detectar amenazas creadas con cualquiera de los vocoders utilizados. A continuación se analizará el rendimiento del sistema al enfrentarse a vocoders desconocidos, especialmente aquellos que realizan un tratamiento realista de la fase.

En el capítulo 5 se probará el rendimiento del sistema ante voces sintéticas adaptadas. Se analizará el funcionamiento del sistema de detección ante voces sintéticas creadas con sistemas TTS que hacen uso de los vocoders utilizados para generar los modelos del sistema. Asimismo, también se hará el estudio con ataques formados por sistemas TTS que hacen uso de los vocoders con tratamiento realista de fase. Por último, se utilizarán señales sintéticas de los sistemas TTS reales más extendidos para evaluar la capacidad del sistema para enfrentarse a amenazas totalmente desconocidas.

El capítulo 6 resume los trabajos llevados a cabo para evaluar diferentes estrategias de entrenamiento de los modelos que se usarán en el sistema SSD. Se comparará el uso del material creado con vocoders con la utilización de material de ataque más realista, y se explorarán las formas de optimización. La base sobre la que se desarrolla esta parte es el ámbito del concurso internacional de detección de voz sintética "ASVspoof 2015: Automatic Speaker Verification Spoofing and Countermeasures Challenge".

Para finalizar, en el capítulo 7 se detallan las conclusiones de la tesis, la difusión de los resultados obtenidos, y los posibles trabajos futuros que quedan abiertos.

2. Estado del arte

2.1. Introducción

La detección de señales sintéticas se enmarca dentro de una aplicación concreta del análisis biométrico: la identificación automática de personas en base a grabaciones de su voz.

Las tecnologías actuales permiten generar de manera relativamente sencilla señales de voz que puedan engañar a los sistemas de identificación de locutor. Este tipo de ataque, denominado spoofing, puede basarse en diferentes métodos: desde algo tan simple como imitaciones o grabaciones, a tecnologías muy elaboradas, como la conversión de voz o síntesis de voces adaptadas. Por ello, se hace necesario desarrollar técnicas que permitan detectar estos ataques, centrándonos en las tecnologías que suponen un reto mayor.

Por otro lado, en el análisis de voz la fase ha sido tradicionalmente dada de lado, por considerar que no contenía información perceptual relevante (Ohm, 1843) (Quatieri, 2002). Sin embargo, cuando se trata de discernir si una señal es humana natural o ha tenido un procesamiento que le haya llevado a tener características de la voz de un hablante objetivo, la fase puede resultar muy útil, dado que la fase de la voz natural tendrá sus características reales, pero en la procesada, probablemente, se haya modificado, si no se ha descartado directamente. Detectar esta diferencia será la base del sistema SSD propuesto en esta tesis.

2.2. Verificación de locutor

Un sistema de verificación automática de locutor (Automatic Speaker Verification, ASV) busca detectar automáticamente a qué persona pertenece la voz de una grabación que se entrega a la entrada. En las últimas décadas se ha hecho muy popular, existiendo sistemas prácticos comerciales. La voz es una de las características más utilizadas en sistemas de acceso biométrico, dado que es una de las más sencillas de obtener de forma no intrusiva.

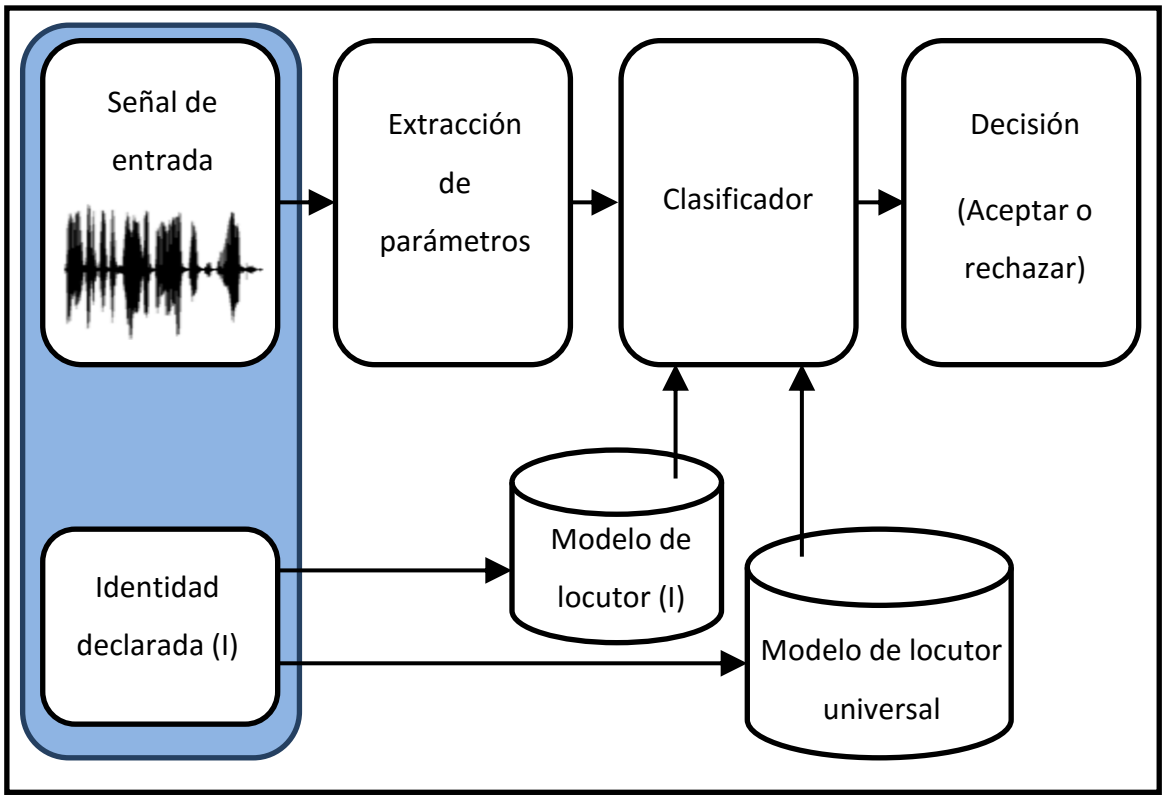

Figura 1: Esquema general de un sistema de verificación de locutor.

En la Figura 1 se puede ver cuál es el esquema básico de un sistema de verificación de locutor según (Wu et al., 2015a). Como entrada del sistema se da una señal de voz, y la identidad que declara quien quiere conseguir el acceso al sistema. De la señal se extraerán los parámetros que el sistema de verificación utilizará para tomar una decisión. Basándose en ellos, decidirá si tienen mayor probabilidad de pertenecer al locutor declarado, o a otro. Para ello el sistema comprobará la verosimilitud de dos modelos: por un lado uno con la información propia del locutor, y por otro lado un modelo universal que recoge muestras de muchos locutores para crear un modelo general que representa a todos los locutores que no son el declarado. En base a ellos se toma la decisión final de aceptar o rechazar la señal de entrada como correspondiente a la identidad que se ha dado.

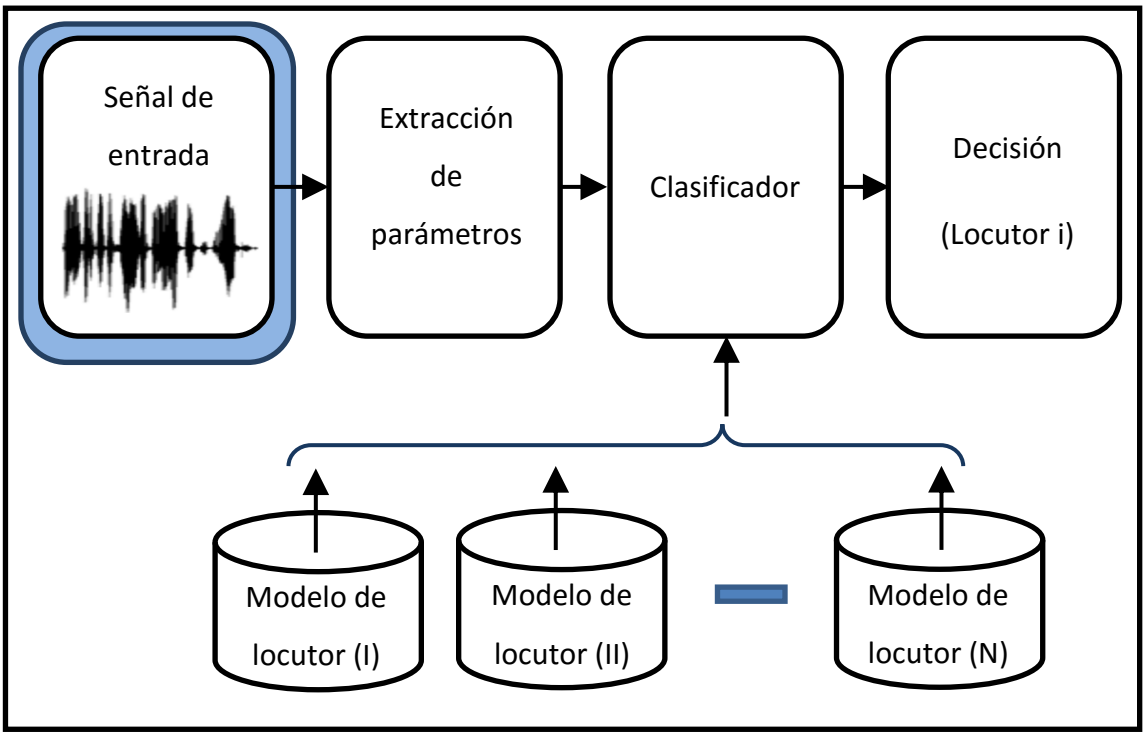

Figura 2: Esquema general de un sistema de identificación de locutor.

Tarea similar es la identificación de locutor (Speaker Identification, SID), que se diferencia de la verificación en que en este caso no hay una entrada al sistema con una declaración de identidad, sino que es el propio sistema el que ha de decidir a qué locutor, de todos los posibles, corresponde una locución de entrada dada, como se ve en la Figura 2, donde tenemos un modelo por cada locutor.

Ambas tareas pueden realizarse de manera dependiente del texto o independiente del texto. En la primera se asume que las frases que se pronunciarán serán fijas o en todo caso, dirigidas por el sistema, y de hecho normalmente suelen ser las mismas durante el entrenamiento del sistema y la verificación. Los sistemas independientes del texto funcionan con frases arbitrarias, incluso idiomas diferentes (Campbell, 1997). La verificación dependiente de texto se ajusta mejor a sistemas de autenticación con frases más cortas, mientras que los sistemas independientes de texto son muy utilizados en entornos de banca telefónica o aplicaciones de vigilancia (Wu et al., 2015a).

Los sistemas de identificación y verificación de locutor han sido referidos en varios trabajos descriptivos, tales como (Rosenberg, 1976), (Campbell, 1997), (Bimbot et al.,

2004), (Faundez-Zanuy y Monte-Moreno, 2005) o (Wu et al., 2015a). En este apartado veremos cómo los distintos módulos han sido resueltos en la literatura.

2.2.1. Extracción de parámetros

Dado que las muestras de la señal de voz no son adecuadas para un tratamiento estadístico, es habitual realizar una parametrización con la que extraer características relevantes. Los parámetros obtenidos se utilizarán tanto en la etapa de modelado y entrenamiento, como en el funcionamiento habitual del sistema.

Los parámetros más utilizados son los espectrales, entre los que podemos encontrar diferentes variantes:

Los parámetros basados en el módulo espectral, y entre ellos los MFCC (Mel Frequency Cepstral Coefficients) (Tokuda et al., 1994) son los más extendidos, habiendo sido utilizados con éxito en muchos ámbitos de las tecnologías de habla. Se han utilizado para verificación de locutor, entre otros, en (Furui, 1981), (Pinto et al., 1995), (Neiberg, 2001), (De Leon et al., 2012) (Bansé et al., 2014) y (Miguel et al., 2014). En (Monte-Moreno et al., 2009) se utilizan realizando una fusión con parámetros LPCC (Linear Prediction Cepstral Coefficients).

Los parámetros frecuenciales LSP (Line Spectral Pair) son muy utilizados en codificación de voz, representando el filtro del modelo de producción vocal. (Campbell, 1997) los utiliza para verificación de locutor.

Entre los parámetros basados en la fase espectral, los RPS representan la fase armónica de una señal. El análisis armónico modela cada trama de una señal como una suma de sinusoides armónicamente relacionadas con la frecuencia fundamental, extrayéndose los parámetros RPS a partir de las diferencias entre las fases de cada sinusoide. Se utilizan para reconocimiento de locutor en (Hernáez et al., 2011). También se han utilizado para verificación de locutor otras representaciones de la fase como MGD (Modified Group Delay) (Wu et al., 2015a)

Recientemente se han empezado a introducir i-vectors (Dehak et al., 2011) para verificación de locutor. A pesar de no ser en sí mismos parámetros para representar una señal de voz como los anteriores mencionados, se pueden utilizar para mejorar la

dimensionalidad y la usabilidad estadística de supervectores creados agregando múltiples parámetros de la señal vocal. Como son teóricamente capaces de modelar las partes más representativas de las características del hablante, se pueden utilizar para verificación de locutor, como en (Bansé et al., 2014).

2.2.2. Creación de modelos

En el esquema general propuesto, una vez seleccionados los parámetros adecuados, es necesario crear con ellos los modelos correspondientes a los locutores que han de ser identificados por el sistema y, en caso necesario, el modelo del locutor universal.

Este modelado se utiliza en el entrenamiento dela mayoría de los sistemas, aunque en algún caso (Furui, 1981) se ejecuta una única realización que se toma de referencia.

El sistema más extendido es la utilización de Modelos de Mezclas Gaussianas (Gaussian Mixture Model, GMM) (Paalanen et al., 2006), tal y como relata (Faundez-Zanuy y Monte-Moreno, 2005). Obtiene resultados particularmente positivos cuando se trabaja con verificación de locutor independiente del texto. Se utiliza este sistema en (Reynolds et al., 2000), (Monte-Moreno et al., 2009), (Campbell et al., 2009), y (Greenberg et al., 2014) entre otros.

Para verificación de locutor dependiente del texto es común también utilizar modelos basados en modelos ocultos de Markov (Hidden Markov Models, HMM) (Bimbot et al., 2004). Este sistema, en verificación independiente del texto, no suele mejorar los resultados obtenidos mediante GMM.

2.2.3. Bases de datos

Los modelos mencionados anteriormente como parte del sistema, tanto los específicos de locutor como los universales, han de generarse para poder ser utilizados, pasando para ello por una fase de entrenamiento. Para ello es necesario disponer de grabaciones de los locutores.

En algunos trabajos se utilizan datos que se recopilan específicamente para cada sistema. En este caso pueden utilizarse grabaciones obtenidas mediante habla microfónica (Furui, 1981) (Pinto et al., 1995) , telefónica (Shaw, 1997) o recopiladas de los medios de comunicación (Fredouille y Charlet, 2014).

Sin embargo, lo más extendido es hacer uso de bases de datos estandarizadas más genéricas, que pueden utilizarse para verificación o para otras tareas.

Las distintas bases de datos NIST (NIST, 1995) son muy utilizadas para verificación de locutor. Desde 1996 se ha venido convocando por parte del NIST el Speaker Recognition Evaluation, cuyo objetivo es impulsar la investigación en reconocimiento de locutor, evaluando el estado del arte en la materia y localizando los algoritmos más prometedores. Para ello, se convoca regularmente una evaluación, en la que ya han participado más de 40 sistemas, entre universidades y sistemas comerciales. Cada evaluación comienza con la publicación del plan, donde, entre otras cosas, se detallan las bases de datos a utilizar, y termina con la presentación en una reunión de los participantes. Entre otros, (Reynolds et al., 2000), (Campbell et al., 2009) y (Novoselov et al., 2014) han generado modelos de verificación utilizando diferentes bases de datos de evaluaciones del NIST.

Gandalf (Melin, 1996) es una base de datos grabada telefónicamente a 86 locutores, y diseñada específicamente para verificación de locutor. Las grabaciones de usuarios legítimos se realizaron en distintas sesiones durante 12 meses, y aporta también grabaciones adicionales de hasta 1000 impostores. Se hace uso de ella en (Neiberg, 2001).

La base de datos YOHO (Campbell y Higgins, 1994), de 186 locutores, está formada por grabaciones de habla microfónica tomadas en ambiente de oficina, en varias sesiones a lo largo de 3 meses. Se utiliza en (Campbell, 1997).

La base de datos RSR2015 (Larcher et al., 2012) está diseñada para verificación de locutor dependiente del texto, recogiendo contraseñas y textos completos de 150 locutores. Se utiliza en (Miguel et al., 2014).

La base de datos Gaudí – AHUMADA (Ortega-Garcia et al., 2000) para identificación de locutor busca reproducir la variabilidad real que se puede encontrar en un sistema, recogiendo grabaciones telefónicas y microfónicas, textos, dígitos y frases fijas, en diferentes sesiones, con 104 locutores. Se utiliza en (Monte-Moreno et al., 2009) entre otros trabajos.

Además de estas bases de datos desarrolladas específicamente para el reconocimiento de locutor, también se ha utilizado en ocasiones para esta tarea la base de datos Wall Street Journal (WSJ) (Paul y Baker, 1992). Ésta, a pesar de estar diseñada para reconocimiento de voz y no de locutor, puede resultar también muy útil para esta tarea al presentar grabaciones de 284 locutores diferentes, con gran calidad y muy poco ruido. Se ha hecho uso de ella, entre otros, en (De Leon et al., 2012) y (Hernáez et al., 2011).

2.2.4. Clasificación

La técnica más extendida para la tarea de clasificación es la obtención de la tasa de verosimilitud: dados una entrada *Y* y un hablante hipotético *S*, trataremos de determinar si *Y* fue dicho por *S*. Partimos de dos hipótesis: que *Y* realmente pertenece a *S* (H0), y que *Y* no pertenece a *S* (H1). Se calculan las verosimilitudes de ambas hipótesis ($p(Y|H0)$, $p(Y|H1)$), y con ellas una tasa de verosimilitud, que suele expresarse de manera logarítmica como diferencia de log-verosimilitudes Λ. La decisión final se toma en base a que dicha diferencia supere o no un umbral dado θ. En la Figura 3 se puede ver el esquema de este tipo de clasificador, según (Bimbot et al., 2004). Este tipo de sistema se aplica en todos los sistemas referidos, salvo los que se mencionan a continuación.

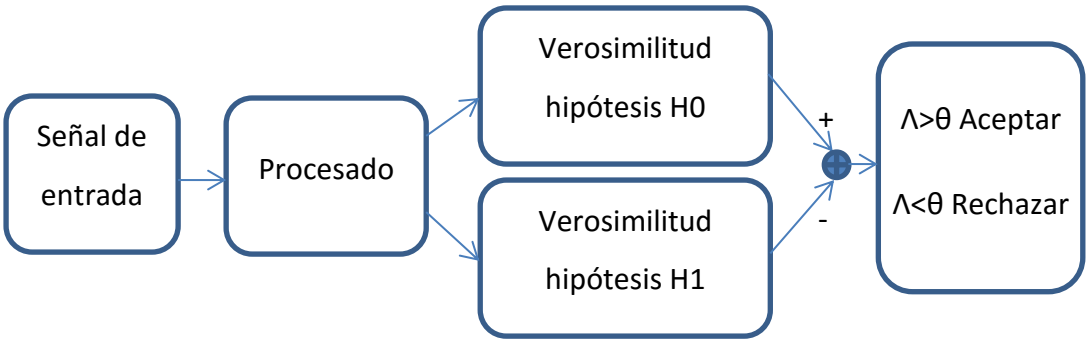

Figura 3: Esquema de un clasificador basado en detección de verosimilitud.

Otro sistema de clasificación que tiene éxito es el basado en redes neuronales artificiales. Éstas se utilizan para modelar un problema imitando el funcionamiento de las neuronas de los organismos vivos: un conjunto de elementos interconectados, sin

una tarea fija para cada uno, pero que durante el entrenamiento van creando y reforzando ciertas conexiones para poder aprender. Se utilizan redes neuronales para el problema de reconocimiento de locutor en (Pinto et al., 1995) y (Campbell, 1997)

También puede enfrentarse el problema de clasificación utilizando redes bayesianas. Son modelos probabilísticos que permiten construir relaciones de probabilidad conjunta que representan las dependencias de un conjunto de variables. Así, aplicado al reconocimiento de locutor, se obtiene una representación de la confianza en que una entrada concreta pertenezca a un determinado locutor. Se proponen redes bayesianas en (Villalba et al., 2013).

2.2.5. Rendimiento de los sistemas de reconocimiento de locutor

Hay diferentes formas de evaluar los resultados, dependiendo principalmente del tipo de sistema, y de la finalidad que se le dé a éste. En el caso de la identificación de locutor solo hay un tipo de error (identificar un locutor diferente al que realmente se presenta) mientras que en la tarea de verificación hay dos tipos de errores posibles: la falsa aceptación (False Aceptance Rate, FAR) cuando se da por buena una identidad que no es correcta y el falso rechazo (False Reject Rate, FRR) cuando se niega el acceso a un locutor legítimo. El sistema puede configurarse para ser más permisivo con una situación o con la otra. Para poder tener un punto de referencia común, en muchas ocasiones se utiliza como medida del error el Equal Error Rate (EER), punto de equilibrio en el que ambos errores, FAR y FRR se igualan. También se puede dar el valor mínimo del DCF (Decision Cost Function), calculado como una combinación lineal del FAR y el FRR, o la precisión (calculada como la tasa de aceptaciones correctas entre las aceptaciones totales) y el *recall* (la tasa entre aceptaciones correctas y candidatos legítimos. También se da en ocasiones la CIR, tasa de identificación correcta.

Por todo ello, diferentes autores expresan sus resultados de formas diferentes, recogiendo a continuación varios ejemplos.

Un sistema antiguo, basado en DTW (Dynamic Time Warping), como el de (Furui, 1981), ya consiguió valores de FAR y FRR de entre el 0,30% y el 9%.

Utilizando Parámetros LSP, (Campbell, 1997) consigue en su mejor resultado una tasa de error de 0,05%.

Los sistemas más extendidos, que utilizan GMM y UBM (Universal Background Model) con MFCC, consiguen valores de error desde el 8% de (Neiberg, 2001) hasta el 2% de (Campbell et al., 2009).

Utilizando i-vectors, como se recopila en (Bansé et al., 2014), se obtienen valores de DCF menores a 0,28. (Fredouille y Charlet, 2014), logra una tasa CIR de hasta el 97,5%, un 9% mejor que la referencia, un sistema equivalente basado en MFCC.

Por último, utilizando parámetros RPS y GMM, (Hernáez et al., 2011) logra un 100% de acierto en la mayoría de los casos de la tarea de identificación, disminuyendo al 98,2% cuando el entrenamiento de los modelos se realiza con solo 5 frases.

A la vista de los resultados expuestos, se puede concluir que la tecnología de verificación de locutor está en un punto de desarrollo avanzado, en el que es posible desarrollar sistemas que sean efectivamente capaces de discriminar la identidad de un locutor con niveles de error bajos, siendo por tanto posible desarrollar sistemas funcionales.

2.3. Spoofing (Simulación)

La verificación de locutor proporciona un acercamiento simple y cómodo a la tarea de autenticación biométrica. Sin embargo, hay una preocupación creciente ante la posibilidad de que un usuario malicioso pueda tratar de engañar al sistema y hacerse pasar por otro, utilizando las denominadas técnicas de *spoofing*: ataques destinados a modificar el resultado del verificador para permitir el acceso a quien legítimamente no debería tenerlo.

Los primeros y más básicos sistemas utilizan grabaciones o imitación de voces para engañar a los sistemas automáticos, habiéndose desarrollado técnicas para neutralizar estos ataques que se detallarán en esta sección. Más recientemente, ha sido el desarrollo de las tecnologías del habla lo que han puesto en jaque la seguridad de los sistemas de verificación, principalmente mediante dos técnicas:

- La conversión de voz (Voice conversión, VC) se basa en modificar las características de la voz de un hablante concreto para, transformándolas, conseguir que suene como otro diferente que será nuestro hablante objetivo.
- La síntesis de voces adaptadas consiste en modificar mediante unas pocas muestras la voz de un sistema TTS para que suene de manera similar a la voz de un hablante objetivo.

Se detallan a continuación los trabajos realizados sobre las distintas técnicas de spoofing mencionadas.

2.3.1. Grabaciones

Un ataque por grabación tiene una estructura muy sencilla: el atacante solo ha de conseguir una grabación de un usuario de un sistema de verificación de identidad por voz, para después presentar al sistema dicha grabación para conseguir acceso. La grabación es un ataque simple, que no requiere conocimientos de tecnologías del habla, y que está ampliamente disponible en una gran cantidad de dispositivos de grabación de bajo coste, como los smartphones. Por tanto, un sistema de detección de ataques de este tipo (PAD, Playback Attack Detection) tiene que ser capaz de descubrir que ha sucedido lo descrito.

2.3.1.1. Técnicas de detección propuestas

En la bibliografía se proponen varias aproximaciones diferentes para conseguir protección frente a las técnicas de grabación.

La aproximación más sencilla y directa para proteger un sistema de verificación de locutor (Speaker Verification, SV) de los ataques de spoofing por grabación es utilizar verificación dependiente del texto: imponer al usuario la necesidad de pronunciar un texto cambiante en cada acceso al sistema (Shang y Stevenson, 2008). Sin embargo, esta implicaría que se perdería la seguridad añadida que puede dar que un cliente pueda dar su contraseña, solamente conocida por él.

Los mismos autores proponen un sistema más elaborado, y que podría funcionar tanto con sistemas SV dependientes del texto, como independientes: almacenar en el sistema las grabaciones de los accesos correctos, de forma que, cuando llega una

nueva solicitud, se puede comprobar si con anterioridad se ha utilizado una señal igual para acceder. Teniendo en cuenta la variabilidad intrínseca del habla, si esto sucede sería señal de haber sido grabada. Se utilizan los picos de la FFT de los sonidos sonoros como parametrización de la voz y para cada intento de acceso al sistema se calcula la similitud con los parámetros de intentos válidos anteriores, denegándose el acceso si dicha medida supera un cierto umbral. Se llegan a obtener unos resultados de EER del 8% con un umbral único para el sistema, y del 4,2% cuando para cada usuario se ajusta un umbral independientemente de los demás.

Otra técnica propuesta se basa en el hecho de que para grabar una señal es necesario un micrófono y para reproducirla un altavoz. Ambos dispositivos pueden dejar huellas mesurables en las señales candidatas del sistema SV. Por ejemplo, el espectro puede quedar aplanado por efecto del ruido y la reverberación de la reproducción. Los altavoces pueden dar una mala respuesta en las bajas frecuencias. Se busca localizar estas distorsiones, denegando el acceso si están presentes. Esta técnica se utiliza en (Villalba y Lleida, 2011a) y (Villalba y Lleida, 2011b), y fue evaluada utilizando la base de datos (NIST, 2010). Se consiguieron valores de EER de entre el 0% y el 7,32% para distintos canales.

Una propuesta parecida que busca distinguir entre el ruido intrínseco que genera el micrófono del sistema y el ruido, diferente, de dos micrófonos (el del sistema, y el que realiza la grabación que se utilizará para atacar) se realiza en consiguiendo reducir la tasa de engaño del sistema de un 40,17% a un 10,26%.

Si se dispone de información adicional también puede utilizarse. Por ejemplo, en un sistema multimodal que contenga audio y video, pueden buscarse incoherencias entre ellas. Así lo propone (Bredin et al., 2006), consiguiendo una tasa de detección del 100% cuando a la grabación de entrada al sistema se le añade una foto fija para el video.

Según lo referido, a priori las grabaciones pueden resultar un medio sencillo y asequible para engañar un sistema de identificación de locutor basado en voz, e indudablemente si no se realiza algún esfuerzo para evitarlo, consiguen engañar a ciertos sistemas de identificación de locutor. Pero se puede luchar contra este tipo de

ataque con modificaciones en la forma en que el usuario accede al sistema, o bien detectando alguno de los rastros que la técnica de grabación dejará, bien sea comparando las grabaciones con otras anteriores legítimas almacenadas en el sistema, bien detectando los dispositivos utilizados para la grabación. Estas medidas de detección consiguen reducir la tasa de falsa aceptación del sistema, pero todavía puede mantenerse relativamente alta, por lo que aún es necesaria más investigación para desarrollar contramedidas verdaderamente eficientes.

2.3.2. Imitaciones

Imitar se define como "Hacer o esforzarse por hacer algo lo mismo que otro o según el estilo de otro". Un imitador sería, consecuentemente, alguien que tiene la capacidad de hacer pasar su voz por la de otra persona. En ese sentido, la imitación de voces es una de las formas más básicas de engaño a un sistema de identificación de locutor, y puede realizarse sin apoyo de ningún tipo de dispositivo hardware o sistema software.

El estudio de las imitaciones de voces se plantea de dos maneras diferentes: mediante el uso de grabaciones realizadas a imitadores profesionales, o mediante imitaciones realizadas por locutores no profesionales.

2.3.2.1. Técnicas propuestas para detectar la imitación.

Los métodos más utilizados para la detección de imitadores de voz se basan en técnicas espectrales, bien mediante el estudio del espectrograma, los formantes, parámetros LFCC (Linear Frequency Cepstral Coefficients) o, mayoritariamente, MFCC. También se ha utilizado la prosodia para detectar imitaciones.

Ya a principios de los 70 en (Endres et al., 1971) se trabajaba en la forma de localizar imitadores de voces, utilizando directamente el espectrograma para identificar cambios forzados en los formantes y otras características visibles.

El análisis de los formantes es la base del sistema propuesto en (Blomberg et al., 2004), en este caso utilizándolos para representar las grabaciones de un locutor profesional e introducirlas en un sistema de verificación de locutor basado en modelos HMM, calculados en base a información obtenida de la base de datos SpeechDat (Elenius y

Lindberg, 1997). El hablante tiene conocimiento de los resultados de reconocimiento del sistema, por lo que puede mejorar su imitación hasta conseguir el acceso.

La tarea de detección de imitadores se plantea mediante un sistema automático basado en MFCC y GMM en (Lau et al., 2004), (Lau et al., 2005), (Farrús et al., 2010) y en (González Hautamäki et al., 2013). Los dos primeros utilizan la base de datos YOHO (Campbell y Higgins, 1994), con imitadores amateur el primero y también con profesionales el segundo. En el tercero se confrontan los resultados de un sistema basado en imitación con los de uno basado en conversión, como los que veremos en 2.3.3. En el cuarto, también se utilizan i-vectors como comparación, y se basa en las bases de datos NIST de los años 2004 (NIST, 2004), 2005 (NIST, 2005), 2006 (NIST, 2006) y 2008 (NIST, 2008).

Por último, también se utilizan GMMs, pero en este caso con parámetros LFCC (Soong et al., 1985), similares a los MFCC pero basados en una escala lineal en lugar de Mel en el sistema de detección de un imitador profesional presentado en (Mariéthoz y Bengio, 2005).

En el experimento relatado en (Zetterholm, 2007) se intenta decidir cuáles son los rasgos más importantes para una buena imitación de voces. Para ello utilizan tres imitadores, dos de ellos profesionales y el tercero amateur. Los tres imitan las voces de nueve reconocidos personajes suecos, tanto políticos como presentadores televisivos. Para tomar una decisión, cada imitación realizada se enfrenta a un experimento subjetivo donde oyentes humanos valorarán la calidad de las imitaciones, y cuáles han sido los rasgos más distintivos para diferenciar. Esta variante de personas detectando imitadores se estudia en profundidad en (Eriksson et al., 2010), donde, desde el punto de vista forense se intenta demostrar qué personas son más adecuadas que otras para la realización de la tarea de detección.

Uno de los parámetros que en (Zetterholm, 2007) resultaron relevantes, la prosodia, se usa también en (Farrús et al., 2008), ya con el concurso de imitadores profesionales intentando engañar a un sistema automático. En este caso se utilizan las voces de dos

imitadores profesionales, que reproducen las voces de conocidos políticos para alimentar un sistema de identificación de locutor que, esta vez, utiliza la prosodia como parámetro para la identificación. Construye para representar las grabaciones, tanto originales como imitadas, un conjunto de parámetros que incluye el ritmo de locución, la duración de los sonidos sordos y sonoros, la media, el valor máximo y mínimo y las pendientes de la frecuencia fundamental, y el jitter y el shimmer.

2.3.2.2. Evaluación de las imitaciones como forma de engañar a un sistema ASV

Se resumen a continuación los resultados obtenidos por los sistemas descritos:

En (Endres et al., 1971) se llegó a la conclusión de que si bien los imitadores conseguían modificar sus formantes, no llegaban a hacerlos coincidir con los de los hablantes imitados.

En (Lau et al., 2004), se llega a concluir que incluso personas sin formación específica en imitación pueden engañar un sistema de verificación de locutor con un pequeño esfuerzo. (Lau et al., 2005) también presenta un sistema automático, basado en GMM y MFCC, aplicado tanto a locutores profesionales como no profesionales. . Los resultados que se obtienen difieren poco entre el grupo de imitadores amateurs y los profesionales en la tarea de hacerse pasar por el locutor más similar de la base de datos, consiguiendo engañar al sistema alrededor del 60% de las ocasiones. Sin embargo, la diferencia aumenta en el caso de otros locutores: en esta situación, los locutores profesionales sí lo hacen mejor.

En las circunstancias propuestas en (Mariéthoz y Bengio, 2005), el sistema de detección automático basado en GMM y LFCC conseguía repeler el ataque sin problemas, no pudiendo el impostor engañar al sistema de verificación ni siquiera en circunstancias de ruido adverso.

En (Zetterholm, 2007) se refuerza que, como cabía esperar, los imitadores profesionales consiguen mejores imitaciones que los amateurs, y que, para los oyentes que realizaron la evaluación, los rasgos más importantes eran la frecuencia fundamental, el dialecto concreto usado, la prosodia, el ritmo y algunos hábitos

fonéticos, pero que, incluso cuando se hace correctamente, en algunos pasajes era posible distinguir la voz del imitador en lugar de la del imitado.

El estudio de parámetros prosódicos realizado en (Farrús et al., 2008) los resultados obtenidos demuestran que el riesgo de engaño del sistema cuando se utiliza exclusivamente la prosodia es alto.

En (Farrús et al., 2010) se confrontan los resultados de un sistema basado en imitación como el anterior, con los de uno basado en conversión, como los que veremos en 2.3.3. Los resultados del primero, con imitadores profesionales tratando de replicar la voz de políticos conocidos, puede llegar a unas tasas de engaño del sistema de alrededor del 20%,

El uso de sistemas automáticos en (González Hautamäki et al., 2013), basados en GMM e i-vectors, llevó a unos resultados que mostraron que el imitador no era capaz de engañar al sistema, habiendo poca diferencia sobre los del sistema original sin imitación.

Como consecuencia de todo lo referido podemos concluir que la imitación de voces, a pesar de que en ocasiones puede ser útil para engañar el juicio humano (y ni siquiera esto durante mucho tiempo), no consigue grandes tasas de éxito en sistemas automáticos complejos. Solamente conformaban una verdadera amenaza en el caso de utilizar para la detección sistemas automáticos muy simples, como por ejemplo los basados exclusivamente en la prosodia, o en el caso de que se pueda acceder a los valores internos de reconocimiento del sistema para un entrenamiento exhaustivo de los imitadores (Blomberg et al., 2004).

2.3.3. Conversión de voces

El objetivo de la conversión de voz es manipular el habla de un locutor dado para que, en algún sentido, se parezca a la de otro, que llamaremos hablante objetivo (Stylianou, 2009), (Erro, 2008), (Evans et al., 2014). A diferencia de los sistemas TTS de síntesis de voz, en los que la entrada al sistema es un texto, la entrada a un sistema de conversión de voz es una grabación natural. Un buen resumen del estado del arte en estas técnicas puede encontrarse en (Wu et al., 2015a).

2.3.3.1. Técnicas de conversión

Habitualmente, la conversión de voz requiere mapeado espectral, relacionado con el timbre de la voz, y conversión de prosodia, relacionado con la frecuencia fundamental, la duración, y otras características prosódicas. Hay tres enfoques para el mapeado espectral: el estadístico-paramétrico, el del desplazamiento de frecuencias (frequency warping) y el de selección de unidades.

El estadístico-paramétrico se basa habitualmente en implementar una función de conversión para mapear las características espectrales de una señal de entrada en aquellas representativas del hablante objetivo. Una aproximación muy directa al problema, haciendo uso de cuantificación vectorial (Vector Quantization, VQ) se presentó en (Abe et al., 1988). Pronto se descubrió que otros sistemas podían realizar la implementación de funciones de conversión de manera más flexible: el uso de GMMs (Kain y Macon, 1998) (Stylianou et al., 1998) (Toda et al., 2007) para obtener funciones lineales, o el de redes neuronales (Desai et al., 2010) o máquinas de Boltzman (Chen et al., 2013) para obtener funciones no lineales.

Las técnicas de frequency warping no buscan sustituir directamente las características de la voz de entrada por las de otra, sino modificar el eje de frecuencias del espectro de la entrada para hacerlo coincidir con el objetivo. De esta manera se mantiene habitualmente un nivel de detalle que con otras técnicas puede perderse. El resultado suele ser una voz convertida de alta calidad (Toda et al., 2001) (Sundermann y Ney, 2003) (Godoy et al., 2012) (Erro et al., 2013). El desplazamiento de frecuencias puede combinarse también con un filtro de desplazamiento de amplitudes (Bonastre et al., 2007).

La conversión basada en selección de unidades parte de un principio similar a la síntesis por selección de unidades: busca en la base de datos de voz del hablante objetivo los segmentos más similares a la señal de voz de entrada (Sündermann et al., 2006) (Dutoit et al., 2007).

Además de la conversión espectral, la prosodia también juega un papel importante en la identificación de las individualidades de los hablantes, principalmente en lo que se refiere a frecuencia fundamental y duración de los sonidos. Para una conversión de

calidad es necesario, por tanto, transformar las trayectorias de pitch de la entrada en las del hablante objetivo (Gillett y King, 2003) (Helander y Nurminen, 2007), así como la duración de fonemas o sílabas (Wu et al., 2006) (Lolive et al., 2008).

2.3.3.2. Uso de conversión de voces para spoofing

Las posibilidades que ofrece la conversión de voz para engañar a un sistema de verificación de locutor han suscitado interés desde hace más de una década. Muchos autores han demostrado la posibilidad de engañar a un sistema ASV basado en GMM-UBM: hasta el 86% de FAR en (Pellom y Hansen, 1999), usando la base de datos YOHO. Un incremento de EER del 16% al 26% en (Perrot et al., 2005), del 8% al 60% en (Matrouf et al., 2006), del 6% al 28% en (Bonastre et al., 2007) y del 8% al 80% en (Alegre et al., 2012), todos ellos usando bases de datos del NIST. Utilizando conversión de voz por selección de unidades, (Kinnunen et al., 2012) muestra un aumento del FAR del 3% a más del 40%.

En general, los estudios realizados sobre el ataque de sistemas SV utilizando conversión de voces han sido más completos que los de imitaciones o grabaciones, y demuestran que esta técnica produce aumentos significativos en las tasas FAR de diferentes sistemas de verificación de locutor.

2.3.4. Síntesis TTS

La síntesis de voz, habitualmente denominada TTS (Text-to-Speech) es una técnica que permite generar una voz natural e inteligible para cualquier texto dado. Es una técnica muy utilizada para distintas aplicaciones, como sistemas de navegación para vehículos, lectores de libros electrónicos, asistencia para discapacitados visuales o ayudas en la comunicación para personas con problemas en el habla.

Los sistemas TTS típicos constan de dos partes, el análisis de texto, y la generación de la forma de onda de la voz. En la primera, se convierte el texto de entrada en elementos lingüísticos (como por ejemplo fonemas) y sus características. En la segunda, se genera la señal de voz a partir de los elementos lingüísticos.

2.3.4.1. Técnicas de síntesis

Existen distintas aproximaciones a la síntesis de voz. A principios de los 70 la generación de la forma de onda se hacía mediante formantes (Klatt, 1980). A partir de los 80, se creaba a partir de una base de datos, relativamente pequeña, de 'difonemas', unidades formadas por la segunda mitad de un fonema y la primera mitad de otro. Éstos se concatenan para generar la forma de onda (Moulines y Charpentier, 1990). En los 90 se optó por recopilar bases de datos de voz más grandes, de las cuales se podían seleccionar fragmentos adecuados en los que tanto el texto como otras características lingüísticas coincidían con las deseadas, consiguiéndose una síntesis de alta calidad con la prosodia adecuada. A este sistema se le denomina genéricamente de 'selección de unidades', y se utiliza en muchos sistemas TTS, como son (Hunt y Black, 1996) (Breen y Jackson, 1998) (Donovan y Eide, 1998) (Beutnagel et al., 1999) (Coorman et al., 2000).

A finales de los 90 surgió otro planteamiento, también basado en bases de datos: la síntesis estadística paramétrica, que se ha hecho muy popular en los últimos años (Yoshimura et al., 1999) (Ling et al., 2006) (Black, 2006) (Zen et al., 2007). En este método, se modelan diferentes parámetros acústicos mediante un sistema estadístico, típicamente basado en modelos ocultos de Markov. Los parámetros acústicos generados por los HMMs se utilizan para alimentar un vocoder que genera la forma de onda de la voz sintetizada.

Por último, recientemente se está investigando el uso de redes neuronales DNN (Deep Neural Networks) para la síntesis paramétrica (Qian et al., 2014).

2.3.4.2. Uso de síntesis para spoofing

Los sistemas más antiguos descritos son en general poco efectivos para utilizar como ataque de spoofing: la síntesis por formantes no es específica de locutor, y los basados en concatenación suelen requerir grandes bases de datos específicas de locutor que cubran, o bien todos los difonemas o grandes cantidades de habla para poder seleccionar el segmento óptimo.

Sin embargo, los sintetizadores paramétricos basados en HMM pueden crear modelos de voz adaptada con una pequeña recopilación de material específico de locutor, adaptado modelos creados anteriormente para otros locutores.

Hay un volumen de información considerable sobre la vulnerabilidad de los sistemas de verificación de locutor ante voces sintéticas. Los primeros trabajos aparecieron hace más de una década: (Masuko et al., 1999) y (Masuko et al., 2000), partiendo de la base de datos en Japonés ATR (Kurematsu et al., 1990) para crear sistemas basados en HMM (tanto para la síntesis como para la verificación de locutor), hicieron que la tasa FAR del sistema de verificación subiera del 7% obtenido con señales exclusivamente humanas, al 70%, utilizando solo 20 locutores.

Se aborda un trabajo más exhaustivo, utilizando la base de datos Wall Street Journal con 283 locutores y dos sistemas de verificación diferentes (uno basado en GMM-UBM y otro en Support Vector Machine, SVM) en (De Leon et al., 2010a), (De Leon et al., 2010b) y (De Leon et al., 2012). Con un sintetizador HMM, la tasa FAR del sistema de verificación subía de 0.28% hasta 86% en el sistema verificador basado en GMM-UBM y de 0% a 81% en el verificador basado en SVM. También (Galou y Chollet, 2011), utilizando sistemas comerciales, llega a parecidas conclusiones.

A la vista de los resultados obtenidos en los sistemas referidos, cuando el habla sintetizada tiene calidad suficiente, es muy posible que, usada como entrada en un sistema de verificación de locutor, sea aceptada como habla legítima del usuario a identificar, incluso en los sistemas de verificación de locutor más avanzados. Esto supone un serio problema de seguridad, y, por tanto, fundamenta la necesidad de desarrollar un módulo de detección de voz sintética que pueda proteger el sistema de verificación de locutor.

2.4. Detección de voz sintética (SSD)

Hasta ahora, se han hecho pocos intentos de discriminar concretamente la voz natural de voz sintética, y tampoco se ha dado con una solución universal (Evans et al., 2013). En general, se utiliza parte del conocimiento sobre el sistema de ataque para poder realizar la detección. Por ejemplo, en algunos trabajos se utilizan ciertas características

de los parámetros espectrales que pueden ser diferentes en ambas versiones (Satoh et al., 2001) (Chen et al., 2010) (Khoury et al., 2014). Otra posibilidad explorada se basa en la dificultad de obtener una síntesis adecuada de la prosodia, de manera que las curvas de pitch se puedan diferenciar específicamente (Ogihara et al., 2005) (Steward et al., 2012). Y la última aproximación busca localizar otras características acústicas que sean diferentes en las señales humanas y en las sintéticas, como es el caso de la fase (De Leon et al., 2011) (De Leon et al., 2012) (Wu et al., 2012) (Wu et al., 2013).

2.4.1. Detección de voz sintética basada en parámetros espectrales

Los parámetros espectrales más utilizados para la tarea de detección de voz sintética son los MFCC. Se utilizan, junto con sus derivadas y la log-energía, en (Satoh et al., 2001) para crear un sistema SSD que funcionará junto con un verificador de locutor, ambos trabajando en paralelo y basados en la base de datos japonesa ATR (Kurematsu et al., 1990). El ataque de un sintetizador de voz paramétrico basado en HMM llevará al sistema de verificación a tasas de falsa aceptación del 86,3%, que pueden reducirse hasta el 0,69% cuando se utiliza el módulo SSD junto con el verificador de locutor.

También se utilizan parámetros Mel Cepstrum, junto con sus derivadas y segundas derivadas, para buscar diferencias entre las componentes de orden más elevado de las señales naturales y sintéticas en (Chen et al., 2010). Aplicándolo a un módulo SSD atacado con un sintetizador de voz paramétrico basado en HMM y STRAIGHT (Speech Transformation and Representation by Adaptive Interpolation of weiGHTed spectrogram), consigue tasas de Equal Error Rate menores al 2%.

De nuevo se utilizan parámetros Mel Cepstrum para detección SSD en (Khoury et al., 2014), creando i-vectors a partir de la base de datos (NIST, 2006), en la que se pueden encontrar ataques de conversión de voz creados para engañar a un sistema de verificación de locutor. Además del que utiliza Mel-cepstrum, también se crea un sistema alternativo basado en LPC (Linear Prediction Coefficients). Ambos fusionan verificación de locutor con SSD, y ambos son capaces de detectar ataque de su propia naturaleza (Mel cepstrum o LPC) con tasas de EER menores al 2%. Sin embargo, el error crece hasta un EER del 53% cuando el método de ataque y el de detección no coinciden.

2.4.2. Detección de voz sintética basada en prosodia

Los parámetros prosódicos, tales como la frecuencia fundamental o la duración de los sonidos, además de servir como información discriminante entre distintos locutores, pueden no ser reproducidos perfectamente por las voces sintéticas, creándose diferencias con las naturales que puedan ser detectadas.

Por ejemplo, la frecuencia fundamental tendrá, habitualmente, más variabilidad en una señal natural que en una sintética, mientras que en esta tendrá mayor posibilidad de formar patrones repetitivos que en una natural. Estas dos características se tratan de detectar en (Ogihara et al., 2005) para detectar señales sintéticas paramétricas basadas en HMM. Utilizando el primero de los criterios se consiguen valores de EER menores al 10% en la tarea de detección, mientras que con la detección de patrones se pueden obtener valores de EER cercanos al 1%.

También se trabaja en la detección de patrones de frecuencia fundamental en (Steward et al., 2012). En este caso, para buscar un escenario de ataques más realistas, se utiliza voz sintética proveniente de diferentes fuentes:

- Síntesis obtenida mediante Festival (Taylor et al., 1998).
- Señales procedentes de los Blizzard Challenge de 2008 (King et al., 2008) y 2011 (King y Karaiskos, 2011).
- Información proveniente de la base de datos WSJ (Paul y Baker, 1992).
- Frases de la TIMIT (Fisher et al., 1986).
- Voces humanas del (NIST, 2002).

Incluso con tan heterogénea información, la búsqueda de patrones de pitch consigue resultados del orden del 10% de Equal Error Rate en la tarea de detección de voz sintética.

2.4.3. Detección de voz sintética basada en la fase armónica

En muchas aplicaciones de las tecnologías del habla aun hoy en día, no se realiza un gran esfuerzo en modelar correctamente la fase, cuando no se descarta directamente, dado que tradicionalmente se ha considerado que el oído humano no la percibía. Esto puede utilizarse como base de un sistema de detección de voz sintética: si en la

generación de voces artificiales la fase no es modelada correctamente, debería ser claramente distinguible de aquella presente en la voz natural.

Recientemente se han propuesto algunas técnicas para aprovechar estas diferencias en la información de fase, utilizando parámetros RPS y MGD.

2.4.3.1. RPS (Relative Phase Shift)

Se utilizan parámetros RPS para desarrollar un sistema SSD en (De Leon et al., 2011), tomando como punto de partida la base de datos WSJ. Los ataques se crean mediante síntesis paramétrica HMM, y el resultado obtenido es de un FAR de 12% y un FRR del 4%. Con la misma base, pero ataques basados en señales transcodificadas mediante STRAIGHT, en (De Leon et al., 2012) se obtiene un EER por debajo del 3% en la tarea de detección SSD.

2.4.3.2. MGD (Modified Group Delay)

La fase del espectro se codifica mediante dos tipos de parámetros en (Wu et al., 2012) usando la base de datos (NIST, 2006). El primero, que denomina *cos-phase* se basa en calcular la transformada discreta de coseno (Discrete Cosine Transform, DCT) de la fase del espectro. El segundo, al que denomina *MGDF-phase (Modified Group Delay Function – phase)*, se basa en la derivada de la fase del espectro, el retardo de grupo, al que también se le aplica la DCT. Con ellos se llega a un rendimiento en la tarea de detección de voz sintética que se evalúa cercano al 4% de Equal Error Rate, muy mejorado respecto de un experimento similar, basado en parámetros de envolvente espectral MFCC, en los que el EER se sitúa en el 15%. El mismo planteamiento se repite, con la base de datos WSJ, en (Wu et al., 2013), consiguiendo reducir el EER de la tarea de detección de voz sintética desde el 10.98% que lograban los parámetros MFCC hasta el 1,25% de MGDF.

Estos resultados hacen ver que es posible crear un módulo de detección de voz sintética basado exclusivamente en la información de la fase armónica de la voz.

2.5.Conclusiones

En los trabajos presentados queda establecido que, si bien la tecnología de verificación de locutor ha llegado a un punto maduro en que, tanto vía microfónica como

telefónica, obtiene unos resultados notables y que permiten su utilización en situaciones reales, las técnicas de spoofing conforman una auténtica amenaza.

Algunas técnicas, como la imitación, no ofrecen un verdadero desafío. En otras, como la grabación, se han desarrollado algunas propuestas que aún necesitan investigación adicional. Pero las técnicas de síntesis de voces adaptadas y de conversión de voces aún pueden utilizarse como manera de engañar a un sistema de verificación de locutor. De ahí que se estén desarrollando formas de detectar específicamente la calidad de sintética de una voz de entrada, en base a diferentes características que la voz puede presentar, como la prosodia, los parámetros de módulo espectral (siendo el máximo exponente de estos la parametrización de envolvente espectral MFCC) o los parámetros de fase.

Esta última, como parametrización RPS, ya se ha probado con éxito en algunos trabajos que han demostrado que la tarea de detección de voz sintética puede realizarse en base a parámetros RPS.

En esta tesis se va a ahondar en esta línea de trabajo, utilizando la parametrización RPS para diseñar un sistema de detección de voz sintética y evaluar su rendimiento en diferentes circunstancias, comparándolo con el rendimiento de los sistemas SSD más habituales basados en parámetros espectrales MFCC.

3. Descripción del sistema SSD

3.1. Introducción

El objetivo de este capítulo es presentar el sistema implementado para ejecutar la tarea de detección de voz sintética.

El sistema busca ser independiente de locutor, de forma que pueda ser utilizado como complemento de un verificador de locutor, pero funcionando de manera independiente, sin necesidad de conocer la información de los locutores registrados en el sistema.

Se busca también que el sistema sea capaz de detectar los ataques generados con las tecnologías más avanzadas disponibles, particularmente los basados en conversión de voz o habla sintética con voces adaptadas. En ambos casos, la gran mayoría de los sistemas actuales hacen uso de vocoders, con los que consiguen parametrizaciones adecuadas para cambiar las características de las voces a modificar, y hacerlas lo más similares posible a las del locutor que se busca suplantar. El sistema buscará la detección de la presencia subyacente de un vocoder en ataque específico, de manera que el éxito en esa tarea nos indicará la presencia de voz convertida o sintetizada que haga uso del vocoder.

La propuesta incluye la utilización de la fase para la detección de voz sintética, mediante la parametrización RPS de las señales a utilizar tanto en el entrenamiento como en la etapa de detección.

El SSD propuesto será un sistema de clasificación basado en GMMs, en el que los parámetros a utilizar serán RPS, y los modelos sintéticos se entrenarán utilizando señales generadas con vocoders. Este sistema tiene la ventaja adicional de no requerir crear voces adaptadas específicas de diferentes locutores, permitiendo que el sistema pueda ser independiente de locutor.

Se tratará de conseguir modelos independientes de locutor en el sistema utilizado, probando con modelos diferentes creados con diferentes locutores, y con diferentes números de locutores.

De cara a validar que, tal como se propone, la información de la fase armónica puede utilizarse como base de un sistema SSD, todos los resultados obtenidos mediante la

parametrización propuesta basada en parámetros RPS serán comparados con los que produzca un sistema canónico típico, que se utilizará como referencia. Éste estará basado en una estructura similar, también clasificando mediante GMM y creando modelos de voces artificiales mediante señales generadas con vocoders, pero la parametrización tendrá un fundamento diferente, basado en parámetros cepstrales de envolvente espectral MFCC (Imai, 1983), que ya se han utilizado en otros trabajos como (Wu et al., 2012) para la detección de voz sintética.

3.2. Diseño del sistema de detección de voz sintética

A continuación se describen las distintas partes del sistema propuesto para la detección de señales sintéticas: el modelo general del sistema, las parametrizaciones utilizadas, el clasificador elegido, los vocoders aplicados, y las bases de datos que proveerán el material vocal necesario.

3.2.1. Arquitectura general

En la Figura 4 se muestra la arquitectura general del sistema SSD implementado.

El sistema de detección es un clasificador binario basado en modelos de mezclas gaussianas. Durante la fase de entrenamiento, se generan GMMs, tanto para la voz natural (λ_{humano}) como para la sintética ($\lambda_{sintético}$).

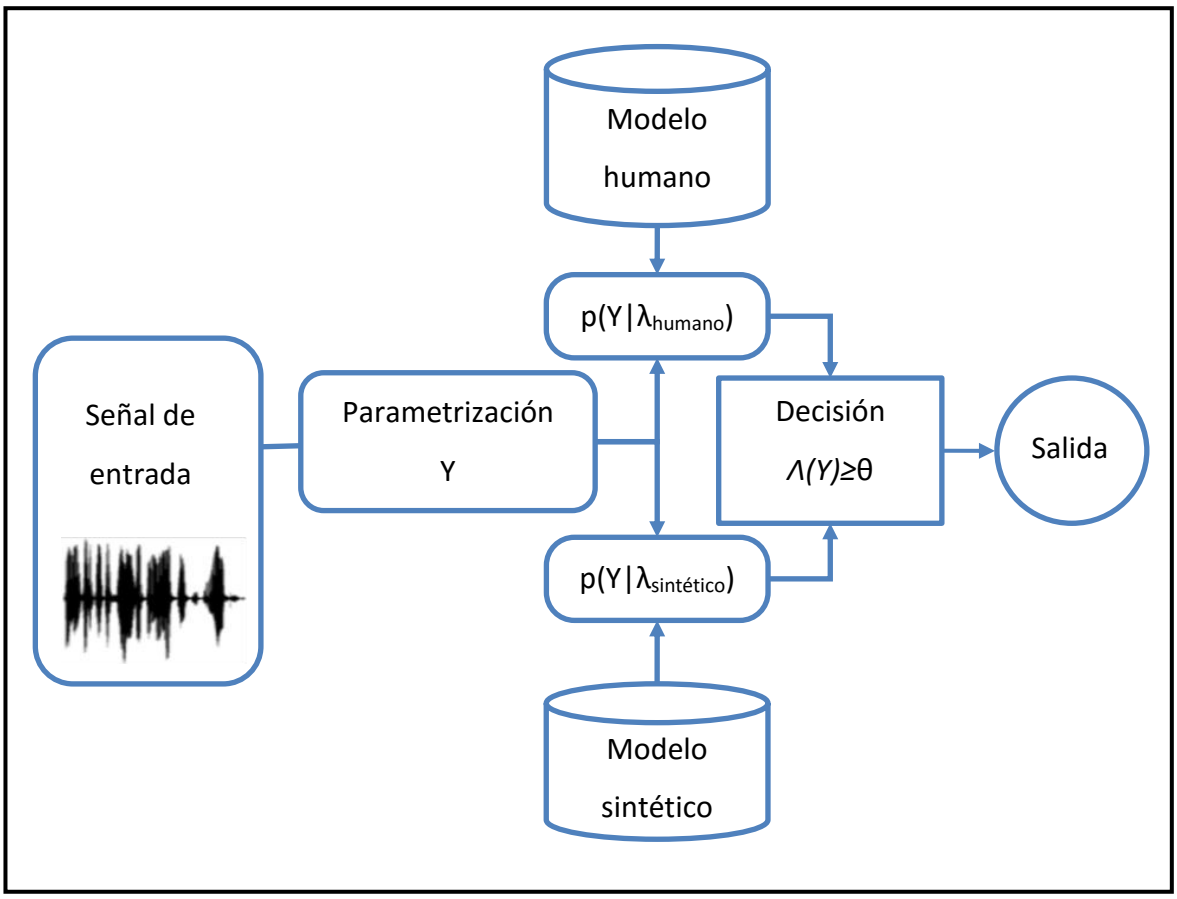

Figura 4: Arquitectura del sistema SSD

Para tomar una decisión sobre si una señal de entrada corresponde a una grabación humana o sintética, el sistema ha de probar la secuencia de vectores de parámetros *Y* de longitud *N* confrontándola contra los modelos natural y sintético, para calcular la verosimilitud correspondiente a cada uno, $p(Y|\lambda_{humano})$ y $p(Y|\lambda_{sintético})$. Entonces se calcula el ratio de verosimilitud Λ según la fórmula de la ecuación (1), tomando como correspondiente a humana la señal de entrada candidata si supera un determinado umbral θ.

$$\Lambda(\mathbf{Y}) = \log p(\mathbf{Y}|\lambda_{humano}) - \log p(\mathbf{Y}|\lambda_{sintético}) \quad (1)$$

donde

$$\log p(\mathbf{Y}|\lambda) = \frac{1}{N}\sum_{n=1}^{N} \log p(\mathbf{y}_n|\lambda) \quad (2)$$

Dependiendo del valor que se asigne al umbral θ, el sistema tendrá mayor tolerancia antes diferentes tipos de errores.

En un sistema de verificación las fuentes de error son dos. Por un lado puede ocurrir un falso rechazo, es decir que se rechace a un locutor válido, que es quien dice ser, y por el otro puede haber falsas aceptaciones cuando se da por válido a un impostor. Estos dos errores dan lugar a sendas medidas del error que comete un sistema que se denominan False Acceptance Rate (FAR) y False Rejection Rate (FRR), calculados según las ecuaciones (3) y (4) respectivamente. El valor del umbral θ hará que el sistema oscile entre uno y otro tipo de error. Así un umbral de decisión bajo hará que el sistema acepte cualquier entrada como válida, incrementando el FAR y minimizando el FRR, y con umbrales de decisión altos ocurrirá al contrario. Por lo tanto, ambos valores de error son necesarios para caracterizar el sistema y definen el punto de operación del mismo.

$$FAR(\theta) = \frac{Número\ de\ candidatos\ impostores\ con\ \Lambda > \theta}{Número\ total\ de\ candidatos} \tag{3}$$

$$FRR(\theta) = \frac{Número\ de\ candidatos\ legítimos\ con\ \Lambda \leq \theta}{Número\ total\ de\ candidatos} \tag{4}$$

Una forma estándar de presentar los resultados de un sistema de este tipo es dibujar el FRR como función del FAR usando para ello diferentes valores del umbral. La curva se representa utilizando una escala basada en la desviación normal, de forma que una distribución gaussiana se vería como una recta en esta escala. De esta forma se obtiene una curva que se conoce como Detection Error Tradeoff (DET), monótonamente decreciente y que permite visualizar fácilmente todos los puntos de operación de un sistema (Martin et al., 1997). En esta curva un punto significativo y que permite resumir el rendimiento de un sistema es el Equal Error Rate (EER) que corresponde con el punto de operación donde los errores de aceptación y rechazo se igualan.

Con un sistema de detección en funcionamiento, y dependiendo del objetivo buscado, es posible buscar diferentes puntos de trabajo según las necesidades. Por ejemplo, si

se desea una seguridad muy alta, con valores muy altos del umbral θ al sistema le costará más aceptar como legítimo un locutor: se darán menos falsas aceptaciones, a costa de aumentar también los falsos rechazos. Para poder comparar diferentes experimentos en un punto común, en todos ellos se ha establecido θ en el punto de Equal Error Rate. Éste punto es aquel en que ambos valores, FAR y FFR coinciden: $EER=FAR(\theta_{EER})=FRR(\theta_{EER})$.

3.2.2. Parametrización

Para poder utilizar la señal de voz con modelos GMM es necesario parametrizarla primero. La parametrización RPS (Saratxaga et al., 2009b) propuesta para el sistema está basada en la fase armónica de la voz. Los resultados obtenidos utilizando RPS se compararán con los obtenidos mediante parametrización MFCC, un sistema canónico típico basado en parámetros cepstrales de módulo (Imai, 1983).

3.2.2.1. Parametrización MFCC

Los parámetros MFCC o Mel Frequency Cepstral Coefficients son coeficientes cepstrales obtenidos a partir del espectro de la señal al que se le ha aplicado previamente un filtrado perceptual basado en la escala mel (Imai, 1983).

Para calcular los parámetros MFCC, se parte de una trama de señal, para la que se calcula su Transformada Discreta de Fourier (DFT), obteniendo su potencia espectral. A ésta se le aplica el banco de filtros correspondiente a la escala Mel (Stevens et al., 1937), calculando el logaritmo de la energía de cada una de las frecuencias Mel. Finalmente, se obtienen los parámetros MFCC aplicando la transformada discreta de coseno a las log-energías extraídas.

Los parámetros MFCC son calculados de esta forma, en nuestro caso concreto cada 10ms en todas las zonas de la señal con actividad vocal. Se obvia el parámetro MFCC-0, de forma que se utilizan 13 parámetros espectrales, junto con sus primeras y segundas derivadas, llevando a vectores de tamaño 39.

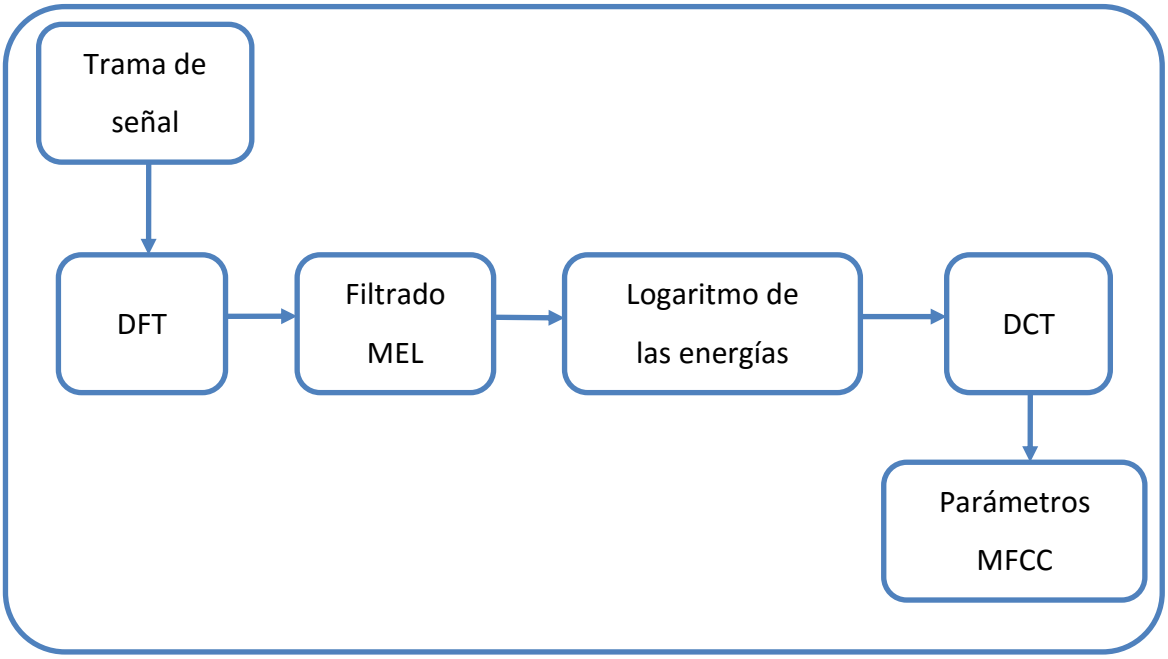

Figura 5: Proceso de parametrización MFCC

3.2.2.2. Parametrización RPS

La representación RPS (Relative Phase Shift) para las fases instantáneas de una señal armónica es la base del sistema de detección propuesto. Ha sido descrita profusamente en (Saratxaga, 2011) y (Saratxaga et al., 2009b).

En estas páginas se va a hacer una descripción del sistema de representación RPS para la fase armónica de la señal de voz.

Definición y derivación

RPS es una representación de la información de fase armónica. El análisis armónico modela cada trama de una señal como una suma de sinusoides armónicamente relacionadas con la frecuencia fundamental, de la siguiente forma:

$$h(t) = \sum_{k=1}^{N} A_k \cos\left(\varphi_k(t)\right) \tag{5}$$

$$\varphi_k(t) = 2\pi k f_o t + \theta_k \tag{6}$$

En estas ecuaciones, N es el número de bandas, A_k la amplitud del k-ésimo armónico, $\varphi_k(t)$ su fase instantánea, f_0 el pitch o frecuencia fundamental y θ_k la fase inicial de la componente k-ésima. La fase instantánea se compone de dos términos:

- La denominada componente lineal, $2\pi k f_0 t$, que depende del instante de análisis y de la frecuencia del armónico.
- La fase inicial de la componente k-ésima, θ_k.

La compleja dependencia de la fase instantánea hace que sea complicada de utilizar para determinadas aplicaciones, como puede ser el análisis de patrones o el modelado estadístico.

Como su propio nombre indica, la representación RPS se basa en calcular la diferencia de fase entre cada armónico y la componente fundamental (k = 1) en un punto específico del periodo fundamental, concretamente el punto donde la fase instantánea del primer armónico es cero, $\theta_1 = 0$.

Los parámetros RPS quedan, por tanto, referidos a un punto concreto fijo dentro del periodo. Esto no implica que el análisis en sí deba realizarse en ese instante específico (por ejemplo, de manera síncrona con el pitch). Por el contrario, si se asume localmente la condición de estacionariedad para la señal, el cálculo de los parámetros RPS puede realizarse en cualquier instante.

Consideremos dos componentes sinusoidales armónicos como los siguientes:

$$x_1(t) = \cos(2\pi f_0 t + \theta_1) \tag{7}$$

$$x_k(t) = \cos(2\pi k f_0 t + \theta_k) \tag{8}$$

En un instante de análisis t_a la fase instantánea de cada componente será:

$$\varphi_1(t_a) = 2\pi f_0 t_a + \theta_1 \tag{9}$$

$$\varphi_k(t_a) = 2\pi f_k t_a + \theta_k \tag{10}$$

Calcularemos los parámetros RPS (Ψ) como la diferencia de fases en el instante t_0, siendo éste el instante más cercano, anterior al punto de análisis, en el que se cumple que $\theta_1(t_0) = 0$. Por tanto, se define:

$$\psi(t_a) = \varphi_k(t_o) - \varphi_1(t_o) = \varphi_k(t_o) \tag{11}$$

Como asumimos la señal como localmente estacionaria, podemos extrapolar el valor de la fase instantánea del k-ésimo armónico en el punto deseado -por ejemplo, $\theta_k(t_0)$-

en base a su valor en t_a. Aceptando que $\varphi_1(t_0) = 0$, podemos obtener t_0 desde la ecuación (6). Obtenemos entonces:

$$t_o = \frac{-\theta_1}{2\pi f_0} \tag{12}$$

A partir de (9) también sabemos que:

$$\theta_1 = \varphi_1(t_a) - 2\pi f_0 t_a \tag{13}$$

Y así, combinando en (11) las ecuaciones (9), (10), (12) y (13), obtenemos:

$$\psi(t_a) = \varphi_k(t_o) = 2\pi k f_0 \left(t_a - \frac{\varphi_1(t_a)}{2\pi f_0} \right) + \theta_k = 2\pi k f_0 t_a + \theta_k - k\varphi_1(t_a) \tag{14}$$

Finalmente, obtenemos la transformación que permite calcular el parámetro RPS para el k-ésimo armónico:

$$\psi_k(t_a) = \varphi_k(t_a) - k\varphi_1(t_a) \tag{15}$$

Representamos en la Figura 6 la interpretación gráfica de la transformación RPS: para el instante de análisis t_a la RPS del armónico k es el desplazamiento de fase de esa componente, respecto del componente fundamental, en el punto en el que comienza el periodo fundamental (t_0).

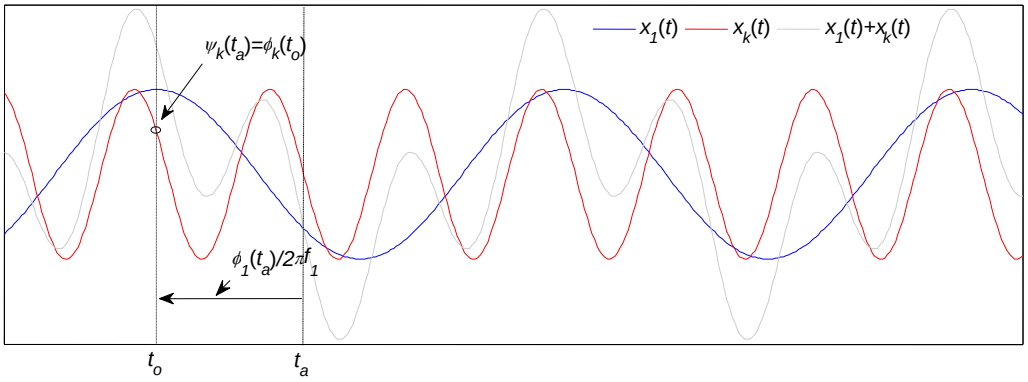

Figura 6: interpretación gráfica de la transformación RPS.

La ecuación (15) define la transformación RPS que permite calcular los parámetros RPS (Ψ_k) a partir de las fases instantáneas en cualquier instante de tiempo de la señal (t_a). Los valores de los RPS se enrollan en el intervalo [-π,π].

Esta representación RPS elimina intrínsecamente el termino lineal de la fase, consiguiendo por tanto una magnitud que se mantiene estable mientras las relaciones de desplazamiento de fase de las componentes no cambien. Estos patrones estables permiten revelar la estructura de fases de la voz, que no es evidente a partir solamente de la fase instantánea. Esto queda de manifiesto en la Figura 7, donde pueden compararse para una señal de voz sonora donde se pronuncian las vocales /aeiou/, los valores de:

a) la fase instantánea
b) la fase expresada como parámetros RPS y
c) la forma de onda.

En las zonas sordas de las señales vocales, al no existir oscilación de las cuerdas vocales, no es posible calcular una frecuencia fundamental. Por tanto, aunque el análisis armónico es posible, ha de realizarse a una frecuencia fija y no desvelará ninguna estructura armónica de fases. Además, en los segmentos sordos la excitación es una señal aleatoria, cuyas fases son consecuentemente aleatorias, no pudiendo por tanto extraer información útil de fase. Por tanto, los fragmentos no sonoros de las señales vocales son descartados en algunas aplicaciones.

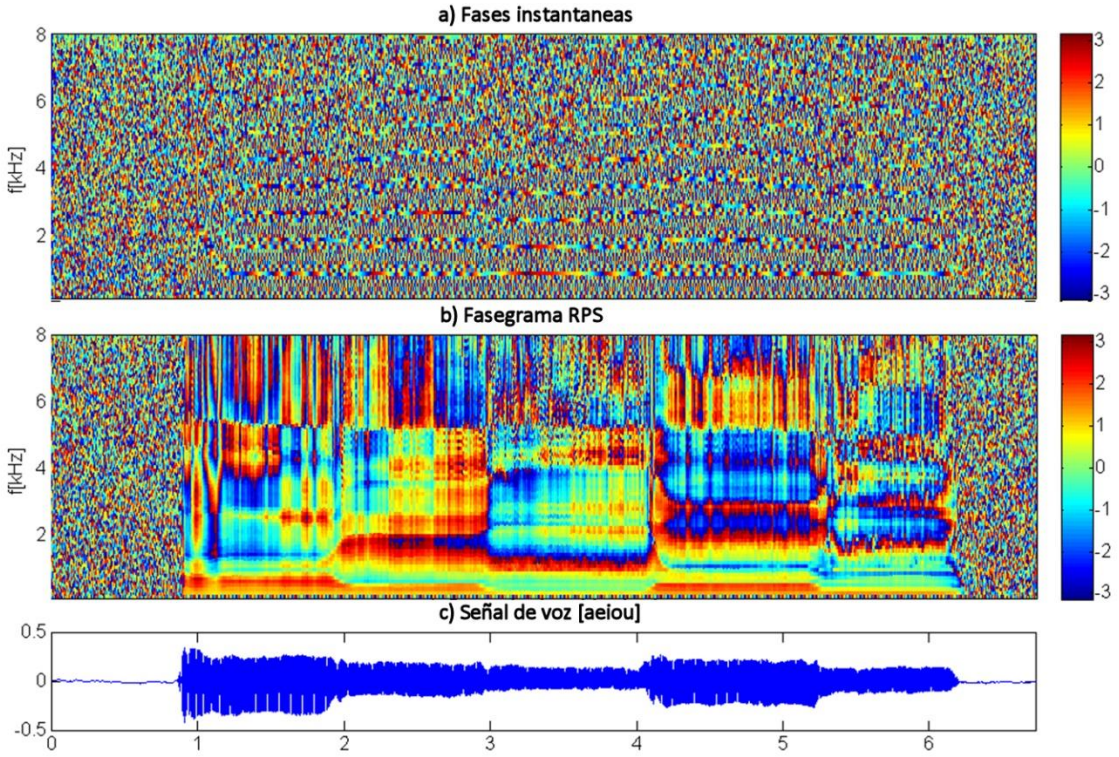

Figura 7: Fasegrama de un segmento sonoro de una señal de voz, con cinco vocales consecutivas /aeiou/.

Obtención de parámetros RPS adecuados para el modelado

Los valores de RPS nos dan una idea clara de la estructura de fases, pero no son adecuados para ser utilizados directamente en un modelado estadístico. El hecho de que la cantidad de parámetros es variable por depender del número de armónicos (que a su vez variará dependiendo del ancho de banda y el valor de la frecuencia fundamental), junto con la alta dimensionalidad y las discontinuidades que puede provocar el desenrollado (unwrapping) de los valores de fase, provocan que sea necesario un procesado adicional para llegar a una parametrización útil para el modelado, la parametrización DCT-mel-RPS que se describe a continuación.

La parametrización DCT-mel-RPS ha sido descrita en profundidad en (Saratxaga et al., 2010) y consigue reducir el número variable de parámetros RPS calculados hasta limitarlos a un número constante.

Para calcular los parámetros, las diferencias de los valores RPS desenrollados son filtrados con un banco de 48 filtros basados en la escala Mel (Stevens et al., 1937)y se aplica una transformada discreta de coseno (DCT) a la secuencia resultante. Dicha DCT se trunca a 20 valores, y se añade un parámetro adicional: el valor promediado de la primera diferencia de los valores RPS desenrollados, la cual, según se demuestra en (Saratxaga, 2011), contiene información relevante.

Para cada vector así obtenido se calculan también sus primeras y segundas derivadas. De esta manera es posible obtener un total de 63 parámetros para cada instante de análisis. En los experimentos de esta tesis, la frecuencia de análisis será de 10 ms, y se aplicará exclusivamente a los fragmentos sonoros.

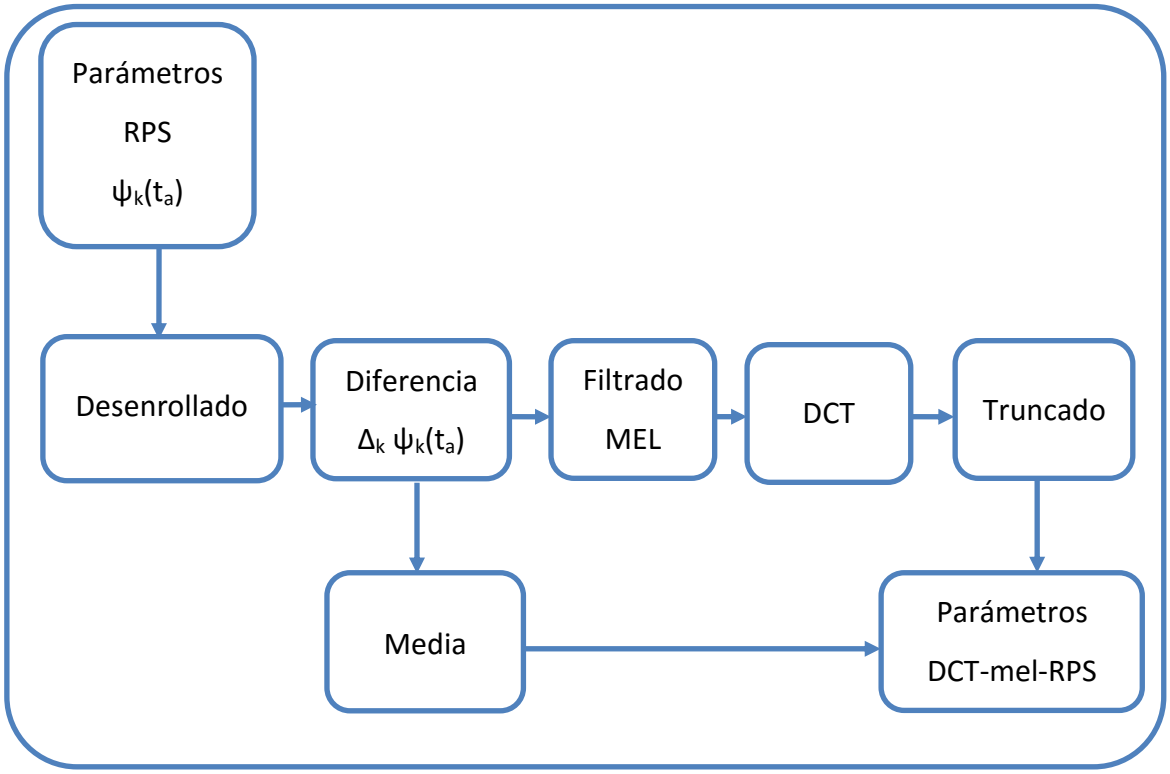

Figura 8: Proceso de parametrización DCT-mel-RPS

Por último, dado que los RPS están directamente relacionados con la forma de onda de las señales, son muy sensibles a un cambio sustancial de la misma, como puede ser el producido por un cambio de polaridad: cuando las ondas de presión en el aire que conforman la voz son recogidas por un micrófono, presiones positivas pueden representarse por tensiones positivas o negativas, dependiendo del conexionado. En

ocasiones ocurre que al realizar una grabación el cableado no es correcto y por tanto la polaridad resulta invertida. Para mantener el rendimiento, a las señales que se parametrizarán mediante RPS se les aplica un detector de polaridad (Saratxaga et al., 2009a), invirtiéndolas en caso de que sea necesario.

3.2.3. Clasificador GMM

El clasificador que utiliza nuestro sistema está basado en Modelos de Mezclas de Gaussianas (Gaussian Mixture Model, GMM). Concretamente, se elaborarán dos clases de modelos, de habla natural y artificial.

En general, los modelos de mezclas de Gaussianas tratan de estimar la función densidad de probabilidad de los parámetros pertenecientes a cada clase *c* mediante una suma ponderada de *M* distribuciones gaussianas (Paalanen et al., 2006):

$$P_c(x) = \sum_{k=1}^{M} \omega_k^c N\left(x, \mu_k^c, \Sigma_k^c\right) \tag{16}$$

siendo $\omega_k{}^c$ el peso de la componente k, con las siguientes condiciones:

$$\omega_k^c > 0 \quad \sum_{k=1}^{M} \omega_k^c = 1 \tag{17}$$

El entrenamiento del modelo consiste en la estimación de los pesos $\omega_k{}^c$ y de los parámetros $\mu_k{}^c$ y $\Sigma_k{}^c$, y se realiza mediante el algoritmo EM (Expectation Maximization). La Figura 9 presenta como ejemplo un GMM en un espacio en dos dimensiones, y cómo es capaz de aproximar una función densidad de probabilidad concreta (Luengo, 2010). Dado un número suficientemente alto de componentes, los GMM permiten aproximar distribuciones continuas, independientemente de su forma.

La clasificación se basa en calcular las probabilidades P(*c*|*x*) de que una muestra *x* pertenezca a cada una de las clases *c*, y seleccionar la más probable. Aplicando la regla de Bayes:

$$\hat{c} = \arg\max_c P(c|x) = \arg\max_c \frac{P(x|c)P(c)}{P(x)} = \arg\max_c P(x|c)P(c) \quad (18)$$

donde P(*c*) es la probabilidad a priori de la clase *c*. En el caso de que la parametrización proporcione Y vectores de parámetros por cada muestra a clasificar, se considera que cada uno de los vectores es independiente de los demás, con lo que la probabilidad conjunta se aproxima por

$$P(x|c) = \prod_{t=1}^{Y} P(x_t|c) \quad (19)$$

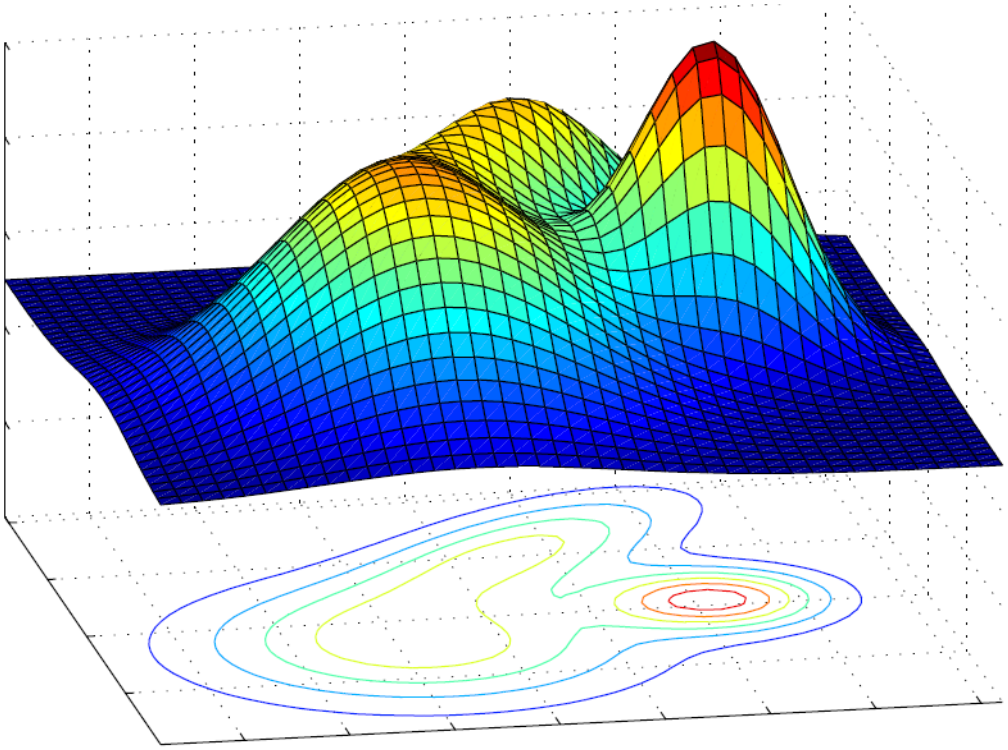

Figura 9: Ejemplo de un GMM en un espacio bidimensional.

Como se ha indicado, con el suficiente número de componentes, un GMM puede aproximar funciones de densidad de probabilidad. Sin embargo, a mayor número de componentes gaussianas se necesitan más datos de entrenamiento para poder estimar sus parámetros. Por tanto, el número de componentes seleccionado implica un compromiso entre la precisión del modelo y su capacidad de generalización: Por un lado, un número de componentes excesivamente bajo, implica que el modelo no podrá aproximar la distribución con suficiente precisión, y estará subentrenado, por lo que

aumentará la probabilidad de error del Sistema. En el extremo contrario, los modelos estarán sobreentrenados, aprendiendo excesivo detalle de la base de datos de entrenamiento, y perdiendo la capacidad de generalizar en muestras desconocidas.

En la mayor parte de la bibliografía se utilizan 512 gaussianas para elaborar los modelos. Siguiendo esa línea, los experimentos relatados en los capítulos 3, 4 y 5 se efectuarán utilizando 512 gaussianas. En el capítulo 6, sin embargo, se realiza un análisis de estrategias para optimizar el resultado obtenido, analizando cuantitativamente el número óptimo de componentes gaussianas.

En la etapa de entrenamiento, los modelos se generaran iterativamente, de modo que se considera que el proceso de creación ha convergido después de un máximo de 10 iteraciones.

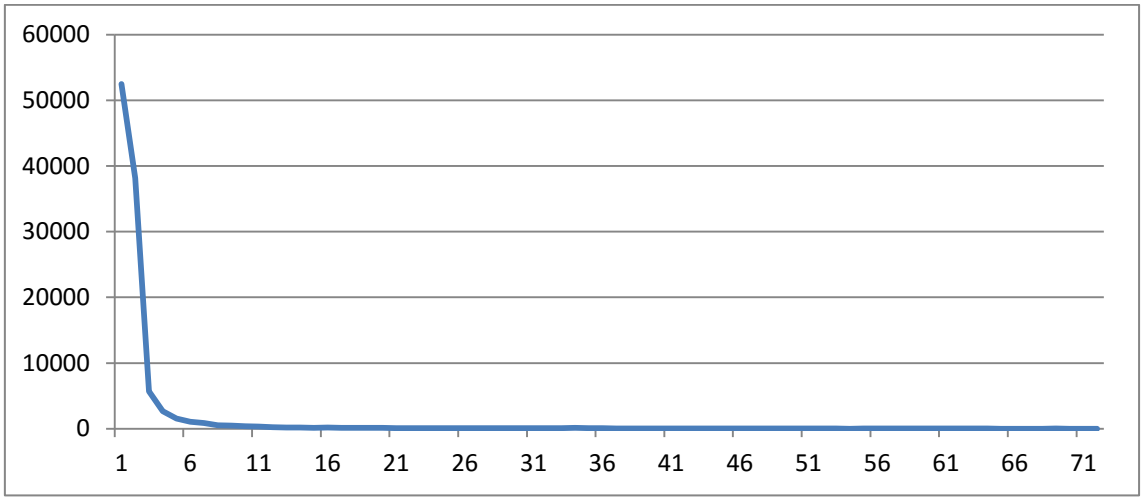

Figura 10: Diferencia de verosimilitud entre iteraciones consecutivas

El motivo de establecer esta condición se ilustra en la Figura 10. En ella se muestran los resultados de un estudio preliminar que se realizó dejando converger libremente la creación del modelo, y cuyo resultado se aprecia: las diferencias significativas en verosimilitud entre iteraciones sucesivas se producen en las 10 primeras iteraciones, manteniéndose muy bajas las diferencias entre iteraciones para las siguientes. Por tanto, dado que, por un lado, reducir el número de iteraciones es positivo por mejorar los tiempos de cálculo de los modelos, y, por otro, se comprueba que las grandes diferencias entre verosimilitudes ya se han producido en las 10 primeras iteraciones,

se acepta como válido limitar el número de iteraciones para la convergencia del modelo a 10.

3.2.4. Modelado

La creación de modelos de voz artificial puede simplificarse si, en lugar de crear un sistema de conversión de voz, o una voz adaptada para cada uno de los locutores del sistema, utilizamos vocoders para generarlas señales sintéticas. En este caso, las señales de voz artificial necesarias para crear los modelos de voz sintética del sistema se han elaborado mediante la técnica denominado copy-synthesis: se codifican utilizando el vocoder deseado, y se vuelve a generar la señal de voz en base a los datos codificados.

3.2.4.1. Descripción de los vocoders utilizados para la generación de señales artificiales de entrenamiento

Se han seleccionado 3 vocoders diferentes, utilizando criterios de popularidad en la tarea de adaptación de voces para síntesis y disponibilidad.

MLSA

Es el vocoder básico incluido en la distribución de demostración del sistema HTS (HTS Working Group, 2002) (Yoshimura et al., 1999). En la etapa de análisis estima la frecuencia fundamental y ejecuta el análisis MFCC (Tokuda et al., 1994) (de orden 24 para una frecuencia de muestreo de 16 kHz). La forma de onda de la señal reconstruida se consigue mediante el filtrado de una excitación de ruido o pulso dependiente de la frecuencia fundamental según sea necesario, a través del denominado filtro MLSA (Mel Log Spectrum Approximation) (Imai, 1983) (SPTK, 2014), relacionado con los coeficientes Mel-cepstrum. Este vocoder ya se ha utilizado para generar señales sintéticas transcodificadas en un sistema de detección (Wu et al., 2012).

STRAIGHT

Este vocoder basado en STRAIGHT también se incluye en la versión de demostración del sistema HTS (HTS Working Group, 2002) (Zen et al., 2007). STRAIGHT es una herramienta de alta calidad para el análisis de voz, manipulación y reconstrucción, que representa una señal vocal en base a su frecuencia fundamental, un envolvente

espectral de alta resolución, y una curva de aperiodicidad de frecuencia (Kawahara et al., 1999). En los sistemas paramétricos de síntesis estadística (Zen et al., 2007), la envolvente espectral está tradicionalmente ligada a la representación Mel-cepstral (de orden 39 para una frecuencia fundamental de 16 kHz), mientras que los valores aperiódicos son promediados en 5 bandas específicas.

El vocoder STRAIGHT es uno de los más ampliamente utilizados para conversión de voces y síntesis, y se ha utilizado como base en el sistema de síntesis utilizado para detección de voz sintética en (De Leon et al., 2012) y (Wu et al., 2013).

AHOCODER

Más reciente que los anteriores, el vocoder AHOCODER (Erro et al., 2011) (Erro et al., 2014) está basado en el modelo armónico con ruido HNM (Harmonic plus Noise Model) (Stylianou, 1996). Parametriza la voz en tres diferentes flujos de datos, como se muestra en la Figura 11:

- Frecuencia fundamental f_0: obtenida mediante un algoritmo preciso de detección de pitch (Luengo et al., 2007).
- Frecuencia máxima sonora (maximum voiced frequency, MVF): obtenida en base a un algoritmo de medida de la similitud sinusoidal (sinusoidal likeness measure, SLM), que indica con qué verosimilitud un pico en el espectro corresponde a una sinusoide. Calculado este parámetro para todos los picos existentes, se obtiene una envolvente SLM de la trama. En base a las frecuencias en que los SLM superan un umbral, se calcula la MVF.
- La envolvente espectral de la señal $S(\omega)$: calculada como coeficientes Mel-cepstrales $\{c_i\}_{i=0\ldots p}$ (de orden 39 para 16 kHz) mediante la siguiente relación:

$$\log S(\omega) = \sum_{i=0}^{p} c_i \cos\left(i \cdot mel(\omega)\right) \tag{20}$$

donde mel(ω) es la versión de ω en escala Mel.

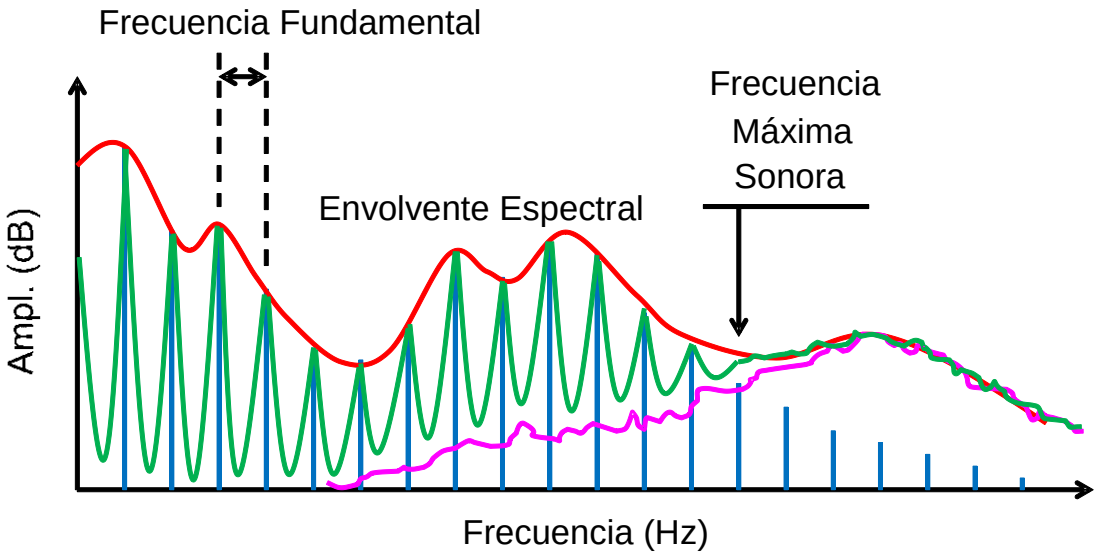

Figura 11: Representación de los tres flujos de datos que forman AHOCODER

La reconstrucción de la señal se realiza utilizando Overlap-and-add (OLA) de las muestras de señal generadas en base a los parámetros para cada trama. Esas muestras se obtienen utilizando procedimientos de HNM: primero se genera la parte ruidosa en el dominio de la frecuencia utilizando el módulo del espectro creado con los parámetros Mel-cepstrales. Para tramas sordas, la señal se obtendrá directamente de la FFT inversa de esta parte. Para tramas sonoras, esta parte ruidosa se pasa por un filtro paso alto dado por la MVF antes de aplicar la FFT inversa. La parte sonora se genera calculando las amplitudes a aplicar en cada uno de los múltiplos de la frecuencia fundamental en base a la envolvente cepstral, y las fases mediante una aproximación de fase mínima que garantiza la coherencia entre tramas contiguas.

El vocoder AHOCODER se ha utilizado para síntesis de voz (Erro et al., 2014) (Erro et al., 2011) y conversión (Erro et al., 2013).

3.2.4.2. Preprocesado

Las señales humanas originales son codificadas y resintetizadas utilizando los tres vocoders para conseguir las señales sintéticas. Antes de la codificación, se filtró la componente continua de las señales, y después, se normalizó la energía de las señales generadas. El motivo de llevar a cabo estas acciones es que no todos los vocoders tratan de la misma forma ni la componente continua, ni la energía. Adicionalmente, la polaridad de las señales fue normalizada, dado que la representación RPS es sensible a

ella. Para esto, se utilizó un algoritmo de detección de polaridad basado en RPS (Saratxaga et al., 2009a), invirtiendo las señales en los casos que fuera necesario para asegurar que todas tuvieran la misma polaridad.

Los tres vocoders utilizados se basan en aproximaciones de fase mínima para la reconstrucción de la fase de las señales vocales transcodificadas. Pero este acercamiento, a pesar de estar muy extendido, no es el único que puede aplicarse. Por ejemplo, otros trabajos como (Drugman et al., 2009), (Raitio et al., 2011b), (Maia et al., 2012) o (Csapo y Nemeth, 2014), proponen métodos de reconstrucción de fase más realistas. Por tanto, se hace necesario un estudio específico sobre este tipo de vocoders, que se llevará a cabo en el capítulo 4.

3.2.5. Bases de datos

Las señales utilizadas para construir los modelos del clasificador, así como para llevar a cabo los primeros experimentos de detección, vienen de la base de datos Wall Street Journal (WSJ) (Paul y Baker, 1992), como en (De Leon et al., 2010b) y (Wu et al., 2013). Las principales ventajas de esta base de datos son el alto número de locutores (283) y la gran relación señal a ruido de sus señales, parámetro importante de cara a conseguir alta calidad en las señales recodificadas con vocoders. Estas características son también las que la han hecho una de las más utilizadas en el ámbito del reconocimiento y verificación de locutor.

La base de datos WSJ viene dividida en varios conjuntos predefinidos. De la misma forma que en (De Leon et al., 2010b), utilizaremos el denominado SI-284, y dentro de éste, un subconjunto de 8599 señales (al que denominamos SSD283) que contiene voces de los 283 locutores. Estas señales se utilizarán para la creación de los modelos humanos.

Las señales humanas se utilizarán también para la creación de señales sintéticas mediante transcodificación utilizando tres vocoders, como se ha detallado en la sección 3.2.4. Todas las señales, tanto originales como sintéticas, se remuestrean a 8kHz, y se parametrizan utilizando parámetros MFCC y RPS como se ha descrito en 3.2.2 para conseguir dos juegos completos de modelos.

Para la evaluación del rendimiento del sistema ante ataques reales y la elaboración de estrategias de modelado se han utilizado también otras bases de datos, que se describirán en profundidad en los capítulos 5 y 6:

- Blizzard 2011 (King y Karaiskos, 2011) y 2012 (King y Karaiskos, 2012): El Blizzard Challenge es una competición donde anualmente se comparan los sintetizadores de voz más punteros disponibles, mediante un corpus común y una evaluación subjetiva a gran escala. Es, por tanto, una muestra representativa de lo más avanzado del estado del arte en síntesis de voz.
- SAS Corpus (Wu, 2015): es una base de datos diseñada específicamente para Spoofing. Comprende voces humanas legítimas grabadas de 45 hombres y 61 mujeres, y voz sintética generada mediante 10 diferentes técnicas de conversión de voces y síntesis.

3.3. Obtención de un sistema independiente de locutor

Cuando se busca un sistema de detección de señal sintética lo más generalista posible, es necesario diseñar el sistema para que sea independiente de locutor sin que la eficiencia se degrade en comparación con su equivalente dependiente de locutor.

Basaremos la independencia de locutor del sistema en diseñar modelos, tanto humano como sintético, que sean independientes de locutor. Para ello, ambos modelos se entrenarán utilizando grabaciones obtenidas de diferentes locutores de forma que al crear los modelos GMM éstos sean capaces de modelar las características comunes a todos ellos y no las diferencias. Estas características comunes deberían serlo también de los locutores cuyas grabaciones no han participado en la creación de los modelos.

Para evaluar el impacto de la cantidad de locutores utilizados para construir los modelos humanos y sintéticos se han definido cinco subconjuntos diferentes de entrenamiento y test a partir de la SSD283:

- 140-SPK: 140 locutores seleccionados aleatoriamente se utilizan para crear los modelos, abarcando aproximadamente 4200 señales. El resto de locutores, 143, son utilizados para el test, con unas 4400 señales aproximadamente entre todos ellos.

- 50-SPK: 50 locutores seleccionados aleatoriamente se utilizan para crear los modelos, abarcando aproximadamente 1500 señales. El resto de locutores, 233, son utilizados para el test, con unas 7100 señales aproximadamente entre todos ellos.
- 30-SPK: 30 locutores seleccionados aleatoriamente se utilizan para crear los modelos, abarcando aproximadamente 900 señales. El resto de locutores, 253, son utilizados para el test, con unas 7700 señales aproximadamente entre todos ellos.
- 15-SPK: 15 locutores seleccionados aleatoriamente se utilizan para crear los modelos, abarcando aproximadamente 450 señales. El resto de locutores, 268, son utilizados para el test, con unas 8150 señales aproximadamente entre todos ellos.
- 5-SPK: 5 locutores seleccionados aleatoriamente se utilizan para crear los modelos, abarcando aproximadamente 150 señales. El resto de locutores, 278, son utilizados para el test, con unas 8450 señales aproximadamente entre todos ellos.

El clasificador se evalúa por separado utilizando señales sintéticas creadas con los tres vocoders utilizados, mencionados en el apartado 3.2.4. Adicionalmente, y para realizar una validación estadística de los resultados obtenidos, cada una de las pruebas se repite cinco veces utilizando diferentes combinaciones de locutores para entrenamiento y para test (con la excepción del juego de locutores 140-SPK, en el que al hacer grupos de 140 locutores sobre un total de 283, con tres repeticiones ya se ha cubierto todo el espectro de locutores posibles).

Como ya se ha descrito en 3.2.1, el umbral del sistema se lleva al punto de Equal Error Rate en cada experimento. Igualmente, para cada una de las repeticiones de la validación cruzada se calcula su propio punto de EER. De esta forma, los resultados se presentarán como la media y la varianza del Equal Error Rate de las repeticiones efectuadas, con el fin de presentar una validación cruzada. Así es como se interpretan en la Tabla 1 y posteriores, donde A se refiere a los valores obtenidos para los

impostores creados con AHOCODER, S a los creados con STRAIGHT y M a los creados con MLSA.

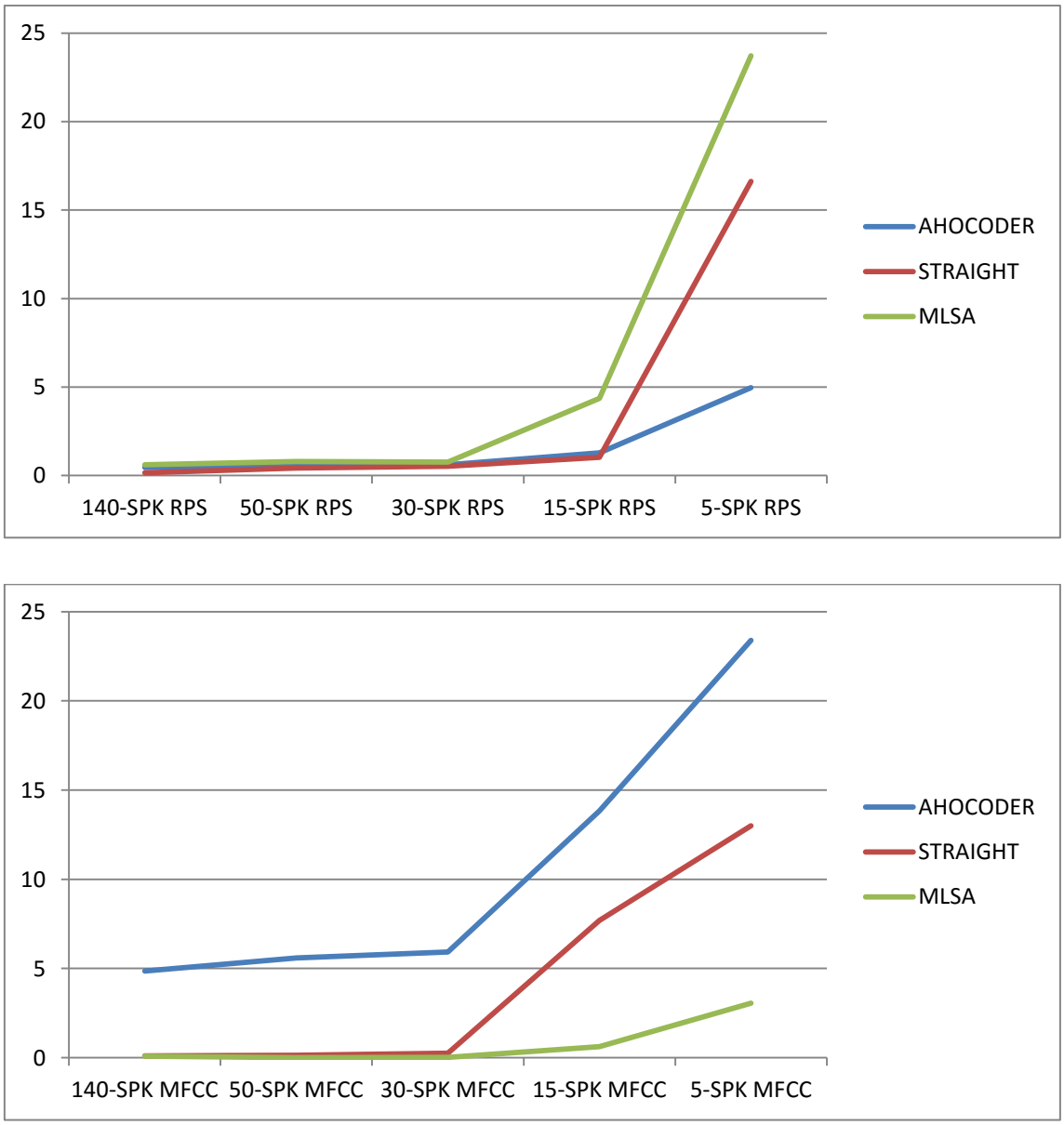

Figura 12: Evolución del EER medio en función del número de locutores utilizado para crear el modelo (arriba, RPS; abajo, MFCC).

	140-SPK		50-SPK		30-SPK		15-SPK		5-SPK	
	RPS	MFCC	RPS	MFCC	RPS	MFCC	RPS	MFCC	RPS	MFCC
m(A)	0,46	4,86	0,51	5,60	0,60	5,93	1,28	13,84	4,97	23,39
σ(A)	0,14	0,67	0,20	0,57	0,16	0,75	0,23	2,12	4,09	3,60
m(S)	0,15	0,10	0,42	0,13	0,52	0,25	1,02	7,69	16,62	13,00
σ(S)	0,09	0,07	0,27	0,08	0,39	0,13	0,22	3,76	41,52	14,85
m(M)	0,61	0,08	0,79	0,01	0,76	0,02	4,35	0,62	23,71	3,06
σ(M)	0,53	0,13	0,31	0,01	0,31	0,01	1,60	0,32	39,89	1,87

Tabla 1: EER medio en porcentaje, y desviación estándar correspondiente, para los diferentes conjuntos de locutores, con impostores creados con AHOCODER, STRAIGHT y MLSA.

En la Tabla 1 y la Figura 12 -que presenta gráficamente los resultados- se observa cómo utilizando ambos tipos de parámetros, RPS y MFCC, se consiguen tasas de error bajas en la tarea de clasificación, siempre que se utilice un número mínimo de 30 locutores para elaborar los modelos. Una vez superado este punto crítico, y aunque este experimento concreto no puede compararse directamente, los resultados se encuentran entre los valores más bajos de EER que se pueden encontrar en la literatura: El sistema SSD basado en parámetros RPS obtiene buenos resultados (por debajo del 1% de EER) de forma consistente para cualquier vocoder. El clasificador basado en MFCC consigue buenos resultados con algunos vocoder (MLSA y STRAIGHT), pero sin embargo tiene una tasa de fallo mayor cuando se enfrenta a señales AHOCODER. En la mayoría de los experimentos se muestra una ligera mejora según se aumenta el número de locutores utilizados para el entrenamiento. La diferencia pasado el punto de 30 locutores, sin embargo, es pequeña, por lo que aceptaremos que puede realizarse un SSD independiente de locutor con cualquiera de los subconjuntos creados con 30 o más locutores. En los experimentos detallados en adelante, los sistemas SSD que se desarrollen se conformarán con el subconjunto 50-SPK.

3.4.Conclusiones

En el capítulo 3 se detallan las características del sistema SSD que se propone utilizar: es un sistema de decisión binaria basado en GMM que utiliza un modelo para las señales naturales y otro para las señales artificiales. Al entrar en el sistema una señal candidata, evalúa su verosimilitud contra uno y otro modelo, y si la diferencia entre ambas supera un cierto umbral, se toma como legítima. Todo ello utilizando parámetros de fase RPS, y comparándolos con parámetros MFCC basados en magnitud que son tomados como baseline.

Para crear el modelo humano se utilizan las señales humanas de la base de datos Wall Street Journal (WSJ), y para el modelo de señales artificiales se utiliza la técnica denominada copy-synthesis: las señales humanas son codificadas por medio de vocoders, concretamente de los tres vocoders AHOCODER, STRAIGHT y MLSA.

Para conseguir que el sistema sea independiente de locutor se crean modelos con señales de múltiples locutores. Se analiza su comportamiento en función del número de locutores utilizados para crear los modelos. Con modelos creados a partir de información proveniente de muy pocos locutores, la capacidad del sistema para obtener resultados de detección fiables es baja. Sin embargo, cuando se utilizan al menos 30 locutores para crear los modelos en los que se basa el sistema, se consigue la capacidad de detectar correctamente la cualidad de sintética. Esto se cumple incluso para señales de locutores no presentes en el modelo, y seleccionando diferentes combinaciones de locutores. En esta tarea, los parámetros RPS basados en la fase armónica de la señal han obtenido, en general, resultados muy similares a la referencia MFCC.

Los resultados de este capítulo se han presentado en (Sanchez et al., 2014).

4. Construyendo un sistema independiente del vocoder

4.1. Introducción

Según se ha visto en el capítulo anterior, un clasificador que utilice un modelo creado con un vocoder específico, puede clasificar correctamente una señal sintética creada con ese mismo vocoder. Pero en un escenario más realista no conoceríamos el tipo de ataque que se está utilizando, ni, si lo hubiera, el vocoder utilizado para crear dicho ataque. Por tanto, necesitamos que el sistema sea independiente del vocoder.

En este capítulo se va a evaluar el rendimiento del sistema cuando se trabaja con modelos provenientes de un vocoder en la tarea de detección de señales creadas con otro vocoder diferente. De esta manera se trata de analizar la dependencia del sistema con el vocoder escogido para crear los modelos.

Posteriormente, se va a buscar la independencia del vocoder construyendo modelos a partir de señales creadas con diferentes vocoders. Con ellos, trataremos de analizar si el hecho de aglutinar en un mismo modelo información proveniente de señales creadas mediante varios vocoders diferentes permite que se generalice mejor, y, por tanto se puedan detectar señales sintéticas obtenidas mediante vocoders que no están presentes en el modelo.

Se creará un sistema final, preparado para detectar ataques basados en cualquiera de los vocoders, utilizando modelos creados a partir de señales transcodificadas con los tres vocoders, y se evaluará su rendimiento en la tarea de detección.

Por otro lado, los vocoders que se han utilizado hasta ahora y que se han descrito en 3.2.4.1 realizan un tratamiento de la fase mediante una aproximación de fase mínima. Sin embargo, se han desarrollado otros vocoders en los que el tratamiento que se hace de la fase lleva a reconstrucciones más realistas. En este capítulo se hará un análisis del rendimiento del sistema propuesto cuando se utilizan dos de estos vocoders que tienen en cuenta el valor de la fase, concretamente GlottHMM y AHOCODER-RPS.

4.2. Análisis de la dependencia del vocoder para el SSD

En el capítulo anterior se ha comprobado que podemos utilizar un modelo de locutor sintético creado con un vocoder específico para detectar señales que también hayan sido creadas utilizando el mismo vocoder. Pero a la hora de tener un sistema funcional

que se enfrente a ataques desconocidos, será imposible saber cuál es el ataque concreto utilizado, e, incluso en el caso de que se hayan utilizado vocoders para generarlo, cuál es el vocoder concreto utilizado para ello. Es necesario por tanto discernir si el sistema SSD tiene la capacidad de clasificar, utilizando modelos basados en un vocoder concreto, señales creadas con otro vocoder diferente.

Con este objetivo se utilizan los modelos del subconjunto SPK-50 descrito en el capítulo anterior. Se crean con información procedente de 50 locutores de la base de datos WSJ, en base a señales transcodificadas con los vocoders AHOCODER, STRAIGHT y MLSA. Cada uno de los modelos se enfrenta a señales de todos los tipos: un juego de cada vocoder, y un último juego de test que incluye todas las señales sintéticas anteriores.

De la misma forma que se realizaba en el apartado 3.3, se utilizan las parametrizaciones MFCC y RPS, y cada prueba es repetida 5 veces con las diferentes combinaciones del subconjunto 50-SPK, de manera que se puede realizar una validación cruzada de los resultados. Para cada combinación de modelo y ataque se calcula independientemente el punto de trabajo, que se ubicará en el de Equal Error Rate. Los valores obtenidos en la Tabla 2 son valores medios del EER obtenido en las distintas repeticiones, y la desviación estándar correspondiente.

Para todas las tablas que aparecen a continuación, en la columna de la izquierda, A representa el vocoder AHOCODER, S representa STRAIGHT, M representa MLSA, y T el último conjunto en el que se incluyen señales de todos los vocoders.

	Modelo AHOCODER		Modelo STRAIGHT		Modelo MLSA	
	RPS	MFCC	RPS	MFCC	RPS	MFCC
m(A)	**0,51**	**5,60**	2,80	25,59	11,17	50,73
σ(A)	**0,20**	**0,57**	0,88	2,44	3,88	1,35
m(S)	2,60	20,27	**0,42**	**0,13**	1,54	13,71
σ(S)	0,88	1,27	**0,27**	**0,08**	0,84	1,18
m(M)	27,04	83,55	26,66	0,93	**0,79**	**0,01**
σ(M)	3,44	3,61	4,55	0,38	**0,31**	**0,01**
m(T)	12,53	36,02	11,80	14,00	7,07	27,46
σ(T)	1,09	0,77	1,48	1,12	2,61	0,62

Tabla 2: EER medio en porcentaje para los experimentos de vocoders cruzados, utilizando parámetros MFCC y RPS.

Los valores que se han resaltado en la Tabla 2 son los obtenidos para la combinación 50-SPK en la Tabla 1. La eficiencia del sistema de detección cuando se trabaja con señales creadas con un vocoder y modelos creados con otro modelo diferente es, en todos los casos, peor que cuando se utiliza para la detección el modelo creado con el mismo vocoder utilizado para generar las señales. Resulta evidente, por tanto, que el rendimiento del sistema es dependiente del vocoder utilizado.

Por otra parte, si se comparan los resultados logrados mediante la parametrización RPS con los que se consiguen utilizando la parametrización MFCC, los números de la Tabla 2 muestran como los resultados de la parametrización de fase son mejores en la mayoría de los casos, para la tarea de detección de vocoders cruzados. Esto sugiere que la parametrización MFCC tiene más dependencia del vocoder que la RPS. Coherentemente con esto, la parametrización RPS da mejores resultados en todos los casos en los que tenemos en cuenta señales creadas con todos los tipos de vocoders disponibles (el set T): 36,02% para MFCC y 12,53% para RPS con el modelo AHOCODER, 14,00% para MFCC y 11,80% para RPS al usar el modelo STRAIGHT, y 27,46% para MFCC y 7,07% para RPS con el modelo MLSA.

4.3. Entrenamiento de un modelo independiente del vocoder

Hasta ahora se ha evaluado la capacidad de detección del sistema utilizando modelos de diferentes vocoders. Los resultados obtenidos han confirmado que cuando el sistema utiliza los modelos de un solo vocoder presenta buen rendimiento para detectar ataques creados con el mismo vocoder, pero obtendrá tasas de error mayores cuando los ataques están generados utilizando vocoders distintos al del modelo. Por ello, y con el objetivo de crear sistemas que puedan obtener buenos resultados para cualquier ataque de entrada, incluyendo aquellos basados en cualquier vocoder, se va a desarrollar a continuación el sistema basado en modelos que incluyen varios vocoders.

Dado que es inviable incluir en un modelo todos los vocoders existentes, se estudiará también la capacidad de extrapolación de los sistemas basados en combinaciones de dos vocoders, evaluando el resultado de detección de señales sintéticas basadas en el vocoder no incluido en el modelo.

El experimento consistirá en crear modelos de múltiples vocoders utilizando señales creadas con varios vocoders, de la misma manera que anteriormente se han creado modelos independientes de locutor incluyendo señales de diferentes locutores. Con ello se persigue un doble objetivo:

- Comprobar que puede agregarse a un modelo señales creadas mediante un nuevo vocoder sin que el rendimiento se degrade gravemente.
- Verificar si los modelos así creados consiguen detectar las señales creadas con un vocoder desconocido mejor que sus modelos equivalentes monovocoder.

De la misma manera que en los experimentos descritos en las secciones 3.3 y 4.2, en los que se trabajaba con modelos creados con información proveniente de un solo vocoder, se han realizado los experimentos siguiendo las siguientes directrices:

- Se utiliza parametrización RPS, pero se desarrolla también un detector basado en parametrización MFCC. Sus resultados serán estudiados como baseline.
- Se utilizan modelos creados con información de 50 locutores, utilizando el subconjunto 50-SPK.

- Para obtener una validación cruzada de los resultados, se repite cada prueba 5 veces, utilizando cada vez juegos diferentes de 50 locutores no solapados elegidos aleatoriamente.
- El umbral de cada sistema se calcula independientemente de los demás, ubicándose en el punto de Equal Error Rate. El valor medio de EER será utilizado para presentar los resultados.

	A+S		S+M		A+M	
	RPS	MFCC	RPS	MFCC	RPS	MFCC
m(A)	0,92	8,67	4,70	41,34	1,12	27,61
σ(A)	0,26	1,30	1,27	2,77	0,23	2,88
m(S)	0,75	0,42	0,70	2,22	1,47	8,81
σ(S)	0,27	0,19	0,38	0,39	0,59	1,24
m(M)	25,32	3,09	0,95	0,03	1,19	0,02
σ(M)	3,07	1,87	0,30	0,01	0,22	0,01
m(T)	11,39	5,42	2,95	22,92	1,32	17,33
σ(T)	1,16	1,07	0,69	1,35	0,38	1,61

Tabla 3: EER medio en porcentaje y desviación estándar para los experimentos con modelos de dos vocoders, utilizando parámetros MFCC y RPS.

Comparando los resultados de la Tabla 3 con los de la Tabla 2, se observa que los modelos multivocoder no son capaces de mejorar los resultados de detección de los modelos monovocoder, cuando se conoce a priori el vocoder concreto utilizado. Por ejemplo, se puede observar que al utilizar el modelo creado con señales correspondientes al vocoder STRAIGHT para detectar señales de esa misma naturaleza, el EER medio obtenido es de un 0,42% para RPS y 0,13% para MFCC, mientras que si se añaden a este modelo señales MLSA (columna STRAIGHT+MLSA) el EER medio crece hasta 0,70% en el caso de la parametrización RPS y hasta 2,20% en el caso de MFCC, incrementos importantes en ambos casos y particularmente en este último, en que llega a ser 17 veces mayor. Es de resaltar la diferencia existente en el ejemplo mencionado, entre los resultados obtenidos por las dos parametrizaciones: mientras

que al utilizar parametrización RPS el error aumenta moderadamente, en el caso de utilizar parámetros MFCC el incremento puede repercutir de manera muy perjudicial en el rendimiento.

En el caso de la tarea de detección basada en modelos de dos vocoders cuando se ataca al sistema con señales basadas en un tercer vocoder diferente, los resultados de la Tabla 3 muestran, al ser comparados con la Tabla 2, que en la mayoría de los casos en que se utiliza la parametrización RPS la introducción de información correspondiente a un nuevo vocoder al modelo resulta beneficiosa (cosa que no sucede en el caso del sistema de referencia basado en parámetros MFCC, donde el hecho de añadir vocoders a un modelo lleva a niveles de detección que se encuentran entre los de los dos vocoders originales). Por ejemplo, al utilizar el modelo creado con STRAIGHT para detectar señales creadas con el vocoder MLSA, el EER medio es de 26,66% para RPS y de 0,93% para la parametrización MFCC. Añadiendo la información del vocoder AHOCODER al modelo, el EER medio mejora hasta llegar a 25,32% en el caso de RPS, mejorando tanto el 26.66% obtenido con el modelo STRAIGHT como el 27,04% obtenido con el modelo AHOCODER. En cambio, el rendimiento con el modelo de dos vocoders AHOCODER y STRAIGHT empeora subiendo hasta el 3,09% para MFCC, quedándose en un punto medio entre AHOCODER (83.55%) y STRAIGHT (0,93%). En ambos casos, aún se está lejos del rendimiento del modelo específico del vocoder MLSA, con un EER medio de 0,79% para RPS y 0,01% para MFCC.

De cara a conseguir un funcionamiento del sistema independiente del vocoder, también se generan modelos utilizando señales creadas con los tres vocoders del sistema, y posteriormente probadas con señales de los tres vocoders contra señales humanas.

	AHOCODER+STRAIGHT+MLSA	
	RPS	MFCC
m(A)	1,16	27,05
σ(A)	0,22	2,30
m(S)	0,99	2,02
σ(S)	0,27	0,35
m(M)	1,32	0,03
σ(M)	0,33	0,02
m(T)	1,22	16,03
σ(T)	0,24	1,79

Tabla 4: EER medio en porcentaje para los experimentos del modelo de tres vocoders, utilizando parámetros MFCC y RPS.

En la Tabla 4 se muestran los resultados correspondientes a los experimentos realizados utilizando información de los tres vocoders para generar los modelos. Utilizando la parametrización RPS se consiguen resultados uniformemente bajos, mientras que con MFCC se llega a una tasa de error muy alta para AHOCODER y excepcionalmente baja para MLSA.

Estos modelos basados en los tres vocoders serán los principalmente utilizados para conformar el sistema de detección independiente del vocoder, como se detallará en los siguientes experimentos.

4.4. Análisis del rendimiento del sistema ante ataques creados con vocoders con tratamiento realista de la fase

Teniendo en cuenta la poca importancia de la fase en la percepción de la señal de voz (Saratxaga et al., 2012), en el ámbito de la síntesis del habla es muy habitual que la información de fase sea directamente desechada, y se genera una nueva fase desde la información relativa a amplitud, usando la condición de fase mínima. Este es el caso, por ejemplo, de los 3 vocoders utilizados en la sección anterior.

Sin embargo, hoy día ya existen algunos vocoders que tratan las fases de manera más realista, y de hecho algunos han sido adaptados a la síntesis de voz estadística (Maia et al., 2007) (Raitio et al., 2011b) (Maia et al., 2012). En los experimentos que se van a describir en este apartado se han probado la validez de modelos desarrollados específicamente con estos nuevos vocoders, así como del modelo de tres vocoders, para detectar señales creadas con vocoders que no utilizan la aproximación de fase mínima.

El objetivo principal de este experimento es buscar los límites de la parametrización basada en RPS: dado que se basa en fase, claramente debe verse afectada por el diferente tratamiento que el vocoder hace a la fase, afrontando de hecho una situación más cercana a la realidad que la que presenta la aproximación de fase mínima. Además, se utilizarán señales de test creadas con estos nuevos vocoders para probar el rendimiento del sistema con el modelo de tres vocoders anteriormente descrito ante ataques desconocidos.

4.4.1. Vocoders con tratamiento realista de la fase

Para elaborar los ataques basados en vocoders con tratamiento realista de fase se han utilizado dos de ellos: GlottHMM y AHOCODER-RPS.

4.4.1.1. GlottHMM

Este vocoder, descrito en (Raitio et al., 2011a) y (Raitio et al., 2011b) y, busca el modelado exacto de la producción de voz humana utilizando filtrado glotal inverso. Con esa técnica se separan y modelan las contribuciones de fuente y tracto vocal a la voz humana.

En el análisis del habla, se utiliza el filtrado inverso iterativo adaptativo (Iterative Adaptive Inverse Filtering, IAIF) (Alku, 1992) en segmentos sonoros para determinar el flujo glotal, que se parametriza mediante varias de sus características:

- La frecuencia fundamental.
- la energía.
- La relación armónico a ruido (harmonic-to-noise ratio, HNR) de cinco bandas.

- El espectro global de la señal de fuente, parametrizada con 10 parámetros LSF (Line Spectral Frequencies).

El tracto vocal estimado utilizando IAIF se parametriza con 30 coeficientes LSF.

En la etapa de síntesis, la fuente de excitación se reconstruye interpolando y escalando un pulso glotal estimado a partir de la voz natural, preservando por tanto la estructura de fase original. La señal de excitación se modifica para reproducir las características variables en el tiempo de la excitación original, y finalmente es filtrada por el tracto vocal para crear la voz.

4.4.1.2. AHOCODER-RPS

Este vocoder se basa AHOCODER, que ya se ha descrito en la sección 3.2.4.1. Sobre él se aplica una modificación diseñada para incluir información sobre las fases originales, utilizando una variante de la parametrización RPS denominada MRRPS (Mel Regularized RPS).

Se basa en la aproximación regularizada discreta cepstrum que se propuso originalmente en (Cappe et al., 1995) para modelar la envolvente espectral utilizando coeficientes cepstrales. Esta aproximación se extendió en (Sorin et al., 2011) para poder modelar fases instantáneas con lo que se vino a llamar representación WMRCC (Weighted Mel Regularized Cepstral Coefficients). Sin embargo, los parámetros así obtenidos siguen sin ser adecuados para su modelado estadístico, de forma que se ha desarrollado un nuevo juego de parámetros que, manteniendo la exactitud de los anteriores, soluciona el problema del uso estadístico.

Utilizando MRRPS se parametrizan las diferenciales de los parámetros RPS desenrollados $\Psi_k'(\omega)$ de una trama acorde a la siguiente expresión:

$$\psi'(\tilde{\omega}) = c_o + 2\sum_{l=1}^{L} c_l \cos(l\tilde{\omega}) \tag{21}$$

$$\tilde{\omega} = G(\omega) \tag{22}$$

Donde los c_j son los parámetros MRRPS y $\tilde{\omega}$ frecuencia normalizada en la escala Mel, en radianes. Se obtiene de la frecuencia lineal en radianes ω mediante la función de distorsión (warping) $G(\omega)$.

El problema reside en determinar el juego de coeficientes tal que la envolvente $\Psi'(\tilde{\omega})$ evaluada en el $\tilde{\omega}_k$ sea lo más parecida posible a los valores originales de RPS diferenciados y desenrollados Ψ_k'. Se impone además una condición adicional para asegurar que la envolvente es suave, representada mediante la función $R[\Psi'(\tilde{\omega})]$ y ponderada por λ, para prevenir condiciones de mal acondicionamiento posterior. Todo ello se resume en un problema de cálculo del error cuadrático mínimo, como se muestra en la siguiente expresión:

$$\varepsilon = \sum_{k=1}^{K} \left\| \psi_k' - \psi'\left(\tilde{\omega}_k\right) \right\|^2 + \lambda R\left[\psi'\left(\tilde{\omega}\right)\right] \tag{23}$$

Según (Cappe et al., 1995) la solución viene dada por la expresión

$$\mathbf{c} = \left(\mathbf{M}^T\mathbf{M} + \lambda\mathbf{R}\right)^{-1}\mathbf{M}^T\boldsymbol{\psi}' \tag{24}$$

donde

$$\mathbf{M} = \begin{pmatrix} 1 & 2\cos\left[G\left(\omega_1\right)\right] & 2\cos\left[2G\left(\omega_1\right)\right] & \cdots & 2\cos\left[LG\left(\omega_1\right)\right] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 2\cos\left[G\left(\omega_K\right)\right] & 2\cos\left[2G\left(\omega_K\right)\right] & \cdots & 2\cos\left[LG\left(\omega_K\right)\right] \end{pmatrix} \tag{25}$$

$$\mathbf{c} = \left[c_0 \cdots c_L\right]^T \tag{26}$$

$$\boldsymbol{\psi}' = \left[\psi_1' \ldots \psi_K'\right]^T \tag{27}$$

$$\underset{i,j=0,\ldots,L}{\mathbf{R}_{i,j}} = 8\pi^2\delta\left[i-j\right]i^2 \tag{28}$$

y λ es el coeficiente de ponderación, cuyo valor se establece en $\lambda = 2\cdot10^{-4}$.

Finalmente, es necesario convertir los L coeficientes MRRPS obtenidos con este modelo en K valores RPS. Para ello, se muestrea en las frecuencias de la escala Mel la

envolvente según la ecuación (21). Finalmente, se integran los valores RPS diferenciados Ψ_k' para obtener los RPS definitivos.

Una vez conseguidos los parámetros de fase de esta forma, la voz se parametriza en 4 flujos diferentes:

- Frecuencia fundamental.
- Coeficientes Mel-cepstrum (de orden 39 para frecuencias de muestreo de 16kHz).
- Frecuencia máxima sonora.
- Los 20 parámetros MRRPS.

4.4.2. Experimentos y resultados

Se utilizan estos dos nuevos vocoders para repetir todo el proceso descrito en el capítulo 3 para los vocoders originales:

- Se resintetizan todas las señales del subconjunto SSD283 de la WSJ utilizando ambos vocoders de fase, para obtener nuevas señales sintéticas.
- Utilizando la distribución 50-SPK, todas esas señales creadas se reparten en los sets de entrenamiento y tests.
- Se generan los modelos sintéticos basados en señales generadas mediante GlottHMM y mediante AHOCODER-RPS.

Una vez generados las nuevas señales y los nuevos modelos, con el sistema creado con ellos se realizan dos pruebas.

En la primera se evalúa la capacidad de discernir entre señales reales y transcodificadas de los modelos entrenados específicamente con la información de los vocoders que hacen uso de la fase: Las señales humanas y transcodificadas mediante GlottHMM se prueban utilizando los modelos humano y GlottHMM, y las señales humanas y transcodificadas mediante AHOCODER-RPS se prueban utilizando los modelos humano y AHOCODER-RPS.

En la segunda, se prueba a detectar las señales creadas con los vocoders nuevos utilizando el modelo de tres vocoders - AHOCODER, STRAIGHT y MLSA - desarrollado en el apartado 4.3.

		GlottHMM		AHOCODER-RPS	
		RPS	MFCC	RPS	MFCC
Modelo específico vocoder de fase	m	0,12	1,19	6,28	8,62
	σ	0.10	0,29	0,53	0,72
Modelo 3 vocoders fase mínima	m	1,37	11,79	3,28	22,17
	σ	0,46	2,27	1,04	2,21

Tabla 5: EER medio en porcentaje para los experimentos con vocoders de fase, utilizando parámetros MFCC y RPS.

En lo que respecta a los modelos específicos de los vocoders que hacen uso de la fase, los resultados muestran que al utilizar la parametrización RPS, los ataques no basados en fase mínima pueden ser correctamente detectados en el caso de GlottHMM. Para AHOCODER-RPS la tasa de error se sitúa en torno al 6%, peor que la mayoría de los resultados anteriores basados en RPS, pero que puede considerarse moderado teniendo en cuenta que el vocoder se ha diseñado específicamente para mantener la estructura RPS original de las señales. Los parámetros MFCC también se ven afectados por el tratamiento de la fase, y se comportan peor que RPS tanto en el caso de señales de test basadas en GlottHMM como en AHOCODER-RPS.

Los resultados obtenidos con el modelo de 3 vocoders y parametrización RPS son muy positivos. Tal y como se observa en la Tabla 5, el modelo basado en RPS muestra buenos resultados, aun a pesar de que no se ha utilizado señales específicas de GlottHMM ni de AHOCODER-RPS para entrenarlos. Específicamente, los resultados de AHOCODER-RPS muestran cómo se detecta mejor con el modelo de 3 vocoders que con su propio modelo específico. Esto da una idea de cómo, cuando el sistema funciona con modelos basados en parámetros RPS, presenta en algunos casos la capacidad de realizar clasificaciones adecuadas de señales producidas mediante vocoders desconocidos.

El rendimiento del sistema en los experimentos realizados con parámetros MFCC es más discreto: su funcionamiento es correcto cuando se utilizan sus modelos específicos para la detección, pero su funcionamiento no mantiene ese nivel cuando se trata de detectar señales sintéticas desconocidas, como demuestran los resultados obtenidos con el modelo de 3 vocoders.

4.5. Conclusiones

Con los experimentos del capítulo 3 se había establecido que la tarea de detección se realizaba correctamente cuando el modelo de voz sintética del sistema de clasificación era el mismo que el utilizado para crear la señal atacante.

En este capítulo se ha estudiado la capacidad del sistema basado en modelos de un solo vocoder para enfrentarse a ataques desconocidos. Concretamente se estudian ataques creados con vocoders diferentes al del modelo. Los resultados en la tarea de detección son sustancialmente peores que cuando se ataca con el vocoder participante en el modelo. Por lo que se concluye que, si no se toman medidas adicionales, el rendimiento del sistema es dependiente del vocoder utilizado en el ataque.

Cuando se extienden los modelos para que representen las características de varios vocoders, se ha comprobado que, en general, esto se sigue cumpliendo. En la mayoría de los casos, al atacar un modelo creado con varios vocoders que no incluyen el vocoder atacante, los resultados obtenidos empeoran frente a utilizar el modelo específico del ataque. Sin embargo, en el caso de la parametrización RPS, la agregación de vocoders en los modelos puede llevar a una mejora de los resultados obtenidos con los modelos de vocoders individuales, cuando el ataque ha sido creado con un vocoder no presente en el modelo. Con RPS, cuando el vocoder utilizado para crear el ataque sí está presente en el modelo, el rendimiento de un modelo multivocoder se degrada un poco frente al del monovocoder correcto, pero manteniendo aún tasas de reconocimiento razonables. Esta característica positiva que se obtiene con la parametrización RPS no aparece cuando se utiliza parametrización MFCC.

De ahí, el crear un modelo con los tres vocoders disponibles, al que llamaremos modelo multivocoder, y que en el caso de vocoders desconocidos puede aprovechar la característica vista de la agregación, y en el caso de los conocidos mantiene una tasa de EER cercana al 1%. Este modelo será el que se utilice en los experimentos siguientes.

En cuanto al estudio del comportamiento del sistema ante vocoders que no utilizan la aproximación de fase mínima, el sistema diseñado para utilizar modelo multivocoder con parametrización RPS es capaz de detectar las señales sintéticas creadas mediante vocoders de fase con un error un orden de magnitud menor que su equivalente de referencia MFCC. Utilizando RPS, lleva a tasas de error cercanas al 1% en el caso de GlottHMM y al 3% en el caso de AHOCODER-RPS. En ambos casos, mejor resultado que el obtenido utilizando ataques y modelos equivalentes generados usando la parametrización MFCC (sobre el 11% y 22% respectivamente).

Los resultados de este capítulo han sido publicados en (Sanchez et al., 2014) y (Sanchez et al., 2015b).

5. Análisis del rendimiento del sistema ante ataques de señales TTS

5.1. Introducción

Para el desarrollo del sistema SSD descrito en capítulos anteriores se han utilizado señales transcodificadas, creadas utilizando el método de copy-synthesis, para ejercer el papel de atacantes o impostores, tanto en la etapa de entrenamiento del sistema como en el test. Se ha podido plantear de esta manera porque se asume que un sistema que utilice estas señales será después capaz de discernir correctamente impostores creados con un método concreto de ataque que haya utilizado ese vocoder para generar la señal, bien sea síntesis o conversión de voz.

De hecho, trabajos previos sugieren que la tarea de detectar señales trascodificadas con un vocoder mediante el método de copy-synthesis es incluso más compleja para un sistema SSD que la detección de una técnica concreta de ataque (De Leon et al., 2012). En todo caso, esta hipótesis, básica para que el sistema diseñado pueda utilizarse en la práctica, debe ser comprobada. Dada la vocación de universalidad del sistema propuesto, es necesario comprobar el rendimiento del sistema cuando se enfrenta a impostores desconocidos, sin restricciones.

El objetivo principal de este capítulo es, por tanto, evaluar la capacidad del sistema para detectar correctamente señales atacantes no procedentes de transcodificación. Entre los métodos más avanzados para conseguir simular la voz de una concreta se encuentran la conversión de voz, y la síntesis de voces adaptadas. Concretamente, en este capítulo se utilizarán como ataques señales sintéticas generadas con diferentes sistemas TTS.

Es interesante recalcar que en estos experimentos se van a utilizar señales de test provenientes de bases de datos diferentes de la WSJ, que ha sido la utilizada para generar los modelos. Por tanto, tanto las condiciones de grabación como, en algunos casos, incluso los idiomas, serán diferentes en las señales usadas para crear los modelos y en las de test. Esta característica acerca las pruebas de rendimiento a una situación real de ataque de spoofing, y permite estudiar el efecto de las diferencias de canal en el sistema de detección.

5.2. Rendimiento del sistema ante ataques creados mediante TTS con vocoder conocido

En este primer experimento se analiza el funcionamiento del sistema de detección cuando se ataca con señales sintéticas en las que, durante su proceso de generación, se han utilizado algunos de los vocoders que también participaban en la fase de entrenamiento. El objetivo es evaluar el rendimiento del sistema que usa modelos creados con señales transcodificadas mediante vocoders, cuando es atacado con señales sintetizadas, provenientes de una base de datos totalmente diferente, pero que hacen uso de los mismos vocoders presentes en los modelos.

5.2.1. Material de evaluación

Se han generado dos nuevos juegos de señales de test para llevar a cabo dos experimentos utilizando dos sistemas diferentes de síntesis estadística:

- Karol: Karol es una voz generada mediante HTS, y entrenada utilizando una base de datos en euskera, sin adaptación de locutor. El vocoder utilizado para parametrizar la voz y reconstruir las señales es AHOCODER (Erro et al., 2014). Para realizar el test, se utilizaron 1000 frases de la base de datos original, se transcodificaron y se sintetizaron utilizando el texto original, utilizando para ello el motor de síntesis estadística de AhoTTS (Aholab, 2011). Para el análisis realizado se utilizaron tanto las señales humanas del locutor original, como las sintéticas generadas con el proceso descrito.
- Karol-RPS: Esta voz es una variante de la anterior. Sigue siendo una voz en euskera, generada mediante HTS, y sin adaptación de locutor, pero en este caso el motor de síntesis ha utilizado el vocoder AHOCODER-RPS descrito en el apartado 4.4: además de la información de AHOCODER, el modelo estadístico también contiene información de fase mediante la inclusión de 20 parámetros MRRPS, junto con sus Δ y ΔΔ. Igual que se ha descrito en el caso anterior, se seleccionaron 1000 frases del locutor original y se sintetizaron utilizando el Motor HTS de AhoTTS, pero en este caso mediante el vocoder AHOCODER-RPS. De nuevo, para el análisis realizado se utilizaron tanto las señales humanas del locutor original, como las sintéticas generadas.

Para realizar las pruebas del sistema SSD, tanto las señales naturales como las sintéticas han sido probadas contra dos modelos:

- El modelo multivocoder utilizado hasta ahora.
- Un modelo monovocoder, creado mediante transcodificación de señales de la base de datos WSJ de manera análoga al anterior, pero incluyendo exclusivamente el mismo vocoder que se ha utilizado para la síntesis: En el caso de Karol, el vocoder es AHOCODER, y en el caso de Karol-RPS, el vocoder es AHOCODER-RPS.

En ambos casos, a pesar de que el vocoder utilizado coincide en el ataque y en el modelo, las señales utilizadas para la generación del modelo no pertenecen a la misma base de datos que las utilizadas para generar los ataques, teniendo características acústicas diferentes.

5.2.2. Resultados

Las tasas de detección obtenidas en estos experimentos se comparan en la Tabla 6 con los que se obtuvieron en el apartado 4.3 para las señales transcodificadas con AHOCODER (Tabla 2 y Tabla 4) y en el 4.4 para las señales transcodificadas con AHOCODER-RPS (Tabla 5).

Modelo \ Evaluación		Karol TTS		AHOCODER		Karol-RPS TTS		AHOCODER-RPS	
		RPS	MFCC	RPS	MFCC	RPS	MFCC	RPS	MFCC
Modelo específico	m	0,12	25,94	0,51	5,60	13,46	23,62	6,28	8,62
	σ	0,04	6,22	0,04	0,34	10,81	5,32	0,53	0,72
Modelo multivocoder fase mínima	m	0,16	56,04	1,16	27,05	57,48	53,64	3,28	22,17
	σ	0,13	6,40	0,05	5,28	28,63	5,96	1,04	2,21

Tabla 6: EER medio en porcentaje para los experimentos con señales sintéticas Karol y Karol-RPS, utilizando parámetros MFCC y RPS.

Los resultados obtenidos, que se muestran en la Tabla 6, son muy diferentes en los dos sistemas de síntesis utilizados. En el caso de Karol, el sistema HTS basado en AHOCODER, el experimento confirma la hipótesis inicial sobre la idoneidad de utilizar señales resintetizadas para entrenar los modelos, al menos en el caso de utilizar parámetros RPS. Se puede comprobar que la detección funciona mejor en el caso de los impostores generados con TTS que para las señales obtenidas mediante transcodificación. Esto se cumple tanto cuando utilizamos para la detección el modelo específico de AHOCODER, como cuando se realizan los test contra el modelo de tres vocoders. Además, las señales con las que se han realizado los tests de señales obtenidas mediante TTS provienen de una base de datos diferente de la original, con canales diferentes e incluso un idioma diferente, lo cual refuerza aún más la hipótesis: es posible elaborar un sistema SSD con modelos creados a partir de transcodificación de vocoder, y obtener buenos resultados aunque las características de las señales de entrenamiento y evaluación sean diferentes.

Los experimentos basados en parametrización MFCC, sin embargo, muestran un comportamiento muy diferente, dado que el resultado obtenido con el test de señales obtenidas mediante TTS es peor que el obtenido mediante transcodificación con el vocoder específico. Comparando estos resultados con los de vocoders cruzados presentados en el apartado 4.2 (donde todos los tipos de ataques se enfrentaban con los tres modelos monovocoder) podemos deducir que los modelos creados mediante la parametrización MFCC son muy dependientes del modelo utilizado, y poco adecuados para extrapolación. Por tanto, en el caso concreto de Karol, las diferencias entre las bases de datos de las que proceden las señales que generaron los modelos y las utilizadas para el test producen un impacto negativo en los índices de detección.

En el caso de Karol-RPS, el sistema de síntesis HTS basado en el vocoder AHOCODER-RPS, los resultados son completamente diferentes. La tasa de detección para las señales de TTS es peor que la referencia obtenida utilizando como ataque señal transcodificada mediante el mismo vocoder, aumentando el error medio del 6,28% al 13,46%. Añadir la información de fase adecuada a la señal sintética es capaz de engañar al sistema SSD tanto en el caso de utilización de parámetros MFCC como

de RPS. La explicación de este fenómeno se puede hallar en el modelado estadístico de la información de fase que se realiza para construir la voz sintética de Karol-RPS: los patrones RPS resultantes son muy diferentes de los utilizados para generar los modelos.

El resultado pobre obtenido en la tarea de detección cuando se usa un sistema basado en los modelos de 3 vocoders de fase mínima sugiere que para conseguir un mejor resultado contra ataques de diversos tipos posibles, los vocoders que incluyen control de fase también deberían ser incluidos en los modelos.

La conclusión de estos resultados es que, cuando se utilizan los parámetros RPS, el uso de vocoders no garantiza unos resultados mejores en la tarea de detección que las señales TTS sintéticas que utilicen esos mismos vocoders, pero, al menos, los resultados muestran que el nivel de error se mantiene bajo. Aun así, la aproximación resulta ser razonable, dado que, por un lado, presenta claras ventajas desde el punto de vista de implementación a la hora de conseguir las señales necesarias para el entrenamiento, y por otro, los sistemas multivocoder RPS tienen capacidad de extrapolación que resulta muy útil a la hora de enfrentarse a escenarios más realistas, tal y como se describe en siguientes apartados.

5.3. Evaluación del rendimiento del sistema ante ataques TTS no restringidos

Dentro de la evaluación del rendimiento del sistema con ataques reales, es necesario evaluar la capacidad de detectar ataque reales basados en TTS con sintetizadores que sean totalmente ajenos a los modelos usados. En este experimento el sistema SSD se utilizará para enfrentarse a múltiples impostores de voz sintética, utilizando el modelo de 3 vocoders, y sin tener conocimiento a priori de las características de las señales con las que se intenta engañar al sistema de detección, en un escenario que puede ser considerado realista. Los atacantes utilizan una tecnología de síntesis desconocida, lo que implica que, aunque en algunos casos efectivamente implica el uso de vocoders para el modelado de las voces artificiales, en otras pueden utilizarse sistemas completamente diferentes. El sistema SSD propuesto se ha diseñado para poder detectar voz sintética generada utilizando síntesis estadística, por lo que es de esperar

que se obtengan mejores resultados detectando ataques de esa naturaleza. Por supuesto, las bases de datos y locutores utilizados para el test en este experimento no son los mismos de las señales de entrenamiento utilizadas para crear los modelos.

5.3.1. Material de evaluación

Con el objetivo de obtener un número representativo de sistemas de síntesis de voz que utilicen la tecnología más actualizada disponible para la tarea de generación del corpus de señales que actuarán como impostoras en la evaluación, se ha optado por utilizar los envíos de los participantes a dos convocatorias recientes del Blizzard Challenge (King, 2014).

El Blizzard Challenge es la convocatoria internacional más popular para la evaluación de sistemas de síntesis de voz. Utilizando como base un corpus oral común, cada participante genera una voz sintética con su sistema TTS. Utilizando esa voz TTS se generan unas señales concretas, que son remitidas a la organización para tomar parte en una prueba subjetiva de percepción. Esta prueba, común para todos los participantes, es llevada a cabo por una gran cantidad de evaluadores.

Los organizadores de este evento científico han conseguido que los sistemas que se presentan a las evaluaciones de los diferentes Blizzard Challenge representen, sin duda, la muestra más avanzada del estado del arte en síntesis de voz.

Para cada edición del Blizzard Challenge, desde la organización de distribuyen los materiales que los diferentes participantes han enviado para su evaluación. Esto incluye, además, un juego completo de señales humanas que pueden ser utilizadas como referencia, y el mismo juego generado por cada uno de los sistemas participantes en la evaluación. De esta manera, este corpus resulta muy útil para poder evaluar un sistema de detección de voz sintética como el aquí presentado.

Las tecnologías de síntesis que toman parte en el Blizzard Challenge son muy diversas, siendo los grupos principales los siguientes:

- Sintetizadores estadísticos basados en HMM: Utilizando un texto de entrada y un modelo de voz estadística, basado en HMM y creado mediante adaptación

desde una voz humana de referencia, se genera una secuencia de parámetros óptimos a partir de la cual se reconstruye la voz.

- Sintetizadores de selección de unidades: Basados en concatenación de forma de onda, dividen la base de datos de entrenamiento en unidades, y a la hora de sintetizar eligen las óptimas según el texto de entrada.
- Sintetizadores híbridos: Generan la señal de voz mediante selección de unidades pero utilizan modelos estadísticos como apoyo para optimizar el propio proceso de selección, modelando estadísticamente alguna de las características que se aplicarán a la voz resultante.

A pesar de que el sistema SSD descrito en este trabajo está específicamente diseñado para enfrentarse a impostores de voz sintética generada mediante vocoders, en el estudio se han incluido todos los participantes en las convocatorias de Blizzard Challenge utilizadas, para que sirvan al menos como referencia de las necesidades futuras de un sistema que pueda ser completamente universal.

En este experimento se han utilizado, concretamente, las señales de evaluación de los Blizzard Challenge 2011 (King y Karaiskos, 2011) y 2012 (King y Karaiskos, 2012).

- Envíos del Blizzard Challenge 2011: En esta entrega se recopilaron 13 juegos de señales, 1 de señales humanas y 12 de señales sintéticas. Como puede verse en la Tabla 7, de ellos, tres (identificados como B, C y D) son sistemas estándar utilizados como referencia, mientras que los otros 9 corresponden a las señales enviadas por los participantes. Respecto de las tecnologías utilizadas, 4 de los sistemas se basan en la selección de unidades (B, E, H y J), otros tantos en síntesis estadística (C, D, F e I), y los 4 restantes pertenecen a la categoría de sistema híbridos (G, K, L y M). Concretamente, el sistema identificado como M es una variación del GlottHMM utilizado para realizar los experimentos de vocoders de fase descritos en el apartado 4.4. En concreto, la implementación del Blizzard 2011 usa el vocoder para la síntesis, pero es alimentado con fragmentos de señales reales. En este experimento se han utilizado las señales de los conjuntos denominados 'novel‘ y ‘news’, que suman 200 señales para

cada sistema. Todas ellas corresponden al mismo locutor y al mismo idioma, inglés de Estados Unidos.

- Envíos del Blizzard Challenge 2012: En este caso, se recopilan 11 juegos de señales, 1 juego humano y 10 sintéticas. De entre las sintéticas, como se recoge en la Tabla 8, uno es un sistema estándar utilizado como referencia (el sistema denominado B), correspondiendo los restantes 9 a los participantes en el Challenge. Concretamente, los sistemas denominados B, F, G e I utilizan la técnica de selección de unidades, mientras que E, H y K son sistemas estadísticos y C y D son híbridos. D en concreto se trata de una actualización del sistema que en Blizzard 2011 se denominaba M, con particularidades similares. Cada juego de señales está formado por 209 locuciones, en inglés de Estados Unidos.

En estos experimentos, cada señal de los conjuntos sintéticos se ha probado junto con su equivalente del juego de señales naturales, de forma que se puede calcular el EER. Y de nuevo, para verificar que el sistema sigue cumpliendo la condición de diseño de independencia de locutor, cada experimento se ha repetido cinco veces, utilizando los cinco modelos multivocoder provenientes del juego 50-SPK descrito en 3.3, de forma que el resultado pueda enfrentarse a validación cruzada.

5.3.2. Resultados

Los resultados de los experimentos realizados con los sistemas participantes en los Blizzard Challenge se muestran en la Tabla 7 y la Tabla 8.

Nombre del sistema	Método utilizado	EER medio con RPS	Desviación con RPS	EER medio con MFCC	Desviación con MFCC
C	Estadística	0,00	0,00	21,80	5,20
D	Estadística	0,00	0,00	12,70	4,01
F	Estadística	0,00	0,00	12,60	2,30
I	Estadística	0,00	0,00	11,56	2,16
B	Selección	51,60	25,24	36,30	3,13
E	Selección	56,60	28,47	51,40	2,22
H	Selección	50,80	26,73	41,00	4,23
J	Selección	59,00	28,02	50,70	3,38
G	Híbrido	53,70	26,13	46,10	3,15
K	Híbrido	52,70	17,28	42,70	4,04
L	Híbrido	60,00	23,27	55,10	1,08
M	Híbrido	33,80	20,81	6,60	1,29

Tabla 7: Tabla resumen de los resultados del experimento realizando con las señales del Blizzard Challenge 2011, utilizando el modelo combinado de los tres vocoders AHOCODER, STRAIGHT y MLSA.

Nombre del sistema	Método utilizado	EER medio con RPS	Desviación con RPS	EER medio con MFCC	Desviación con MFCC
E	Estadística	0,00	0,00	34,64	7,71
H	Estadística	0,00	0,00	60,67	14,35
K	Estadística	11,19	10,63	0,00	0,00
B	Selección	44,78	1,57	39,42	4,30
F	Selección	76,84	4,05	86,22	4,39
G	Selección	28,52	10,98	95,21	1,51
I	Selección	40,57	1,80	41,34	4,19
J	Híbrido	70,05	11,99	11,77	2,18
C	Híbrido	55,69	0,99	62,30	4,62
D	Híbrido	52,73	8,48	51,87	4,86

Tabla 8: Tabla resumen de los resultados del experimento realizando con las señales del Blizzard Challenge 2012, utilizando el modelo combinado de los tres vocoders AHOCODER, STRAIGHT y MLSA.

Los resultados conseguidos con los modelos elaborados con parametrización RPS son excelentes: utilizando el modelo de 3 vocoders se consiguen distinguir correctamente los impostores con una exactitud total, esto es, con un EER del 0% en todos los casos de síntesis estadística excepto uno (el sistema K de Blizzard 2012, que sí ha alcanzado el 0% de error con MFCC). Estos resultados avalan la capacidad de extrapolación del sistema desarrollado con modelos basados en RPS, dado que utilizando exclusivamente señales resintetizadas se ha conseguido entrenar un modelo capaz de detectar correctamente señales sintéticas de naturaleza totalmente desconocida. Este no es el caso, sin embargo, de los modelos basados en MFCC, los cuales han conseguido, con pocas excepciones, resultados más pobres. Por ejemplo, en los sistemas estadísticos de la edición de 2011, el EER de los sistemas basados en MFCC está entre el 11% y el 21%, mientras que se mantenía a cero para los sistemas basados en RPS.

Respecto de los sintetizadores basados en selección de unidades e híbridos, el sistema SSD propuesto no ha sido capaz de detectar las señales sintéticas generadas con ellos. Esto entra dentro de lo que cabía esperar, dado que el detector se ha diseñado para filtrar las señales sintéticas generadas utilizado vocoders, pues estas son las que permiten la adaptación de voces necesaria para imitar a un hablante específico. Obviamente, los sistemas de selección de unidades e híbridos generar otro tipo de distorsiones tanto en la fase de la señal como en el módulo, que deberían ser incluidas en los modelos. Este es el caso de los sistemas M de 2011 y D de 2012, variaciones de GlottHMM, donde el modelo específico de vocoder ha mostrado un funcionamiento mejor. En todo caso, los sistemas híbridos y de selección de unidades no están diseñados para poder ser aplicados directamente a la adaptación de voces objetivo, y por tanto es necesario evaluar hasta qué punto desempeñan una amenaza real para los sistemas de acceso biométrico, quedando mientras tanto fuera del alcance de este trabajo.

5.4. Conclusiones

En este capítulo, el sistema SSD diseñado se ha puesto a prueba con distintas señales generadas mediante sintetizadores de voces adaptadas.

En las primeras pruebas, realizadas con la voz denominada Karol basada en AHOCODER, y su equivalente Karol-RPS basada en el vocoder de fase AHOCODER-RPS, podemos concluir que la detección de vocoders no produce necesariamente mejores resultados en la detección de voces sintéticas que en la detección de los propios vocoders. Pero el error se mantiene en niveles bajos, lo que unido a las ventajas desde el punto de vista de la obtención de señales para el entrenamiento de modelos y la capacidad de extrapolación de los sistemas multivocoder RPS, demuestran que la aproximación presentada, que usa modelos de vocoders es razonable. Los resultados pobres obtenidos con las señales atacantes basadas en Karol-RPS indican que es necesaria plantear trabajos posteriores para poder mejorar la detección de los vocoders que hacen un uso realista de la fase, por ejemplo, incluyendo de alguna manera información de dichos vocoders en los modelos.

Las pruebas realizadas con los sintetizadores presentados al Blizzard Challenge, que en su momento son los más avanzados disponibles, muestran que el sistema basado en el modelo de tres vocoders RPS es capaz de detectar sin error casi todos los sintetizadores basados en modelos estadísticos, que son los más adecuados para generar voces específicas, y por tanto la mayor amenaza para los sistemas de verificación de locutor.

Los sistemas basados en selección de unidades no han podido ser correctamente detectados, si bien es necesario evaluar el nivel de riesgo que suponen realmente desde el punto de vista de la suplantación. En general, la elaboración de sistemas de síntesis basados en selección de unidades que tengan una calidad suficiente requiere de la grabación de grandes bases de datos del locutor al que representará la síntesis, y estas son habitualmente complicadas de obtener en un escenario en el que se intenta atacar un sistema de acceso y no se cuenta con la colaboración de la persona suplantada.

Los resultados de este capítulo se han publicado en (Sanchez et al., 2015b), artículo que ha recibido el premio al mejor artículo 2015 de la Red Temática de Tecnologías del Habla.

6. Estrategias de entrenamiento del sistema SSD

6.1. Introducción

En los capítulos anteriores se ha descrito el sistema desarrollado utilizando modelos creados en base a señales de la base de datos WSJ a los que se aplica la técnica de copy-synthesis utilizando diversos vocoders, y el rendimiento que se obtenía usando ese diseño. El motivo principal para generar los modelos artificiales con señales basadas en transcodificación mediante vocoders era la dificultad de conseguir los sistemas adaptados necesarios, bien sea mediante sistemas de conversión de voz o mediante síntesis adaptada, para conseguir ataques realistas que además tuvieran calidad suficiente para poder engañar a un sistema de verificación de locutor.

En este capítulo se revisará la hipótesis del uso ventajoso de vocoders para la creación de modelos de voz sintética para usar en sistemas SSD, utilizando una base de datos específicamente creada para evaluar la tarea de detección de voz sintética utilizando ataques reales, ASVSpoof2005 (Wu et al., 2015b), (Wu et al., 2015c). La base de datos está formada por una serie de señales que forman una representación completa de los ataques reales punteros que pueden representar una amenaza, según la tecnología del momento, para los sistemas SSD. Esto incluye síntesis de voz adaptada utilizando sistemas TTS, y ataques basados en conversión de voces. Éstos no han sido aún probados como ataque en nuestro sistema SSD.

Utilizaremos esta nueva base de datos para estudiar diversas estrategias de entrenamiento de modelos. En una primera propuesta, diseñaremos una estrategia para mejorar los resultados que se pueden obtener con los modelos descritos en capítulos anteriores, basados en la base de datos WSJ, utilizando los datos de entrenamiento de la nueva base de datos. En una segunda propuesta, se utilizarán exclusivamente los datos provenientes de ASVSpoof2005 para aplicar la estrategia de copy-synthesis mediante vocoders y crear los nuevos modelos sintéticos. En ambos casos se compararán con el rendimiento obtenido usando modelos creados exclusivamente con los ataques reales. Se utilizarán las métricas de evaluación diseñadas para el *Automatic Speaker Verification Spoofing and Countermeasures Challenge* para evaluar el rendimiento de los sistemas entrenados en base a las estrategias diseñadas.

6.2.La base de datos de ataques reales ASVSpoof2015

Tras la sesión especial sobre spoofing y contramedidas en el ámbito de la verificación de locutor que en año 2013 tuvo lugar en Lyon dentro de Interspeech (Evans et al., 2013), y siguiendo con esa visión que buscaba dar a conocer la seriedad del spoofing e impulsar la investigación en ese ámbito formando una comunidad que colabore en la creación de basases de datos y estándares específicos, en 2015 se presenta el desafío "ASVspoof 2015: Automatic Speaker Verification Spoofing and Countermeasures Challenge". En él, se desarrolla una base de datos estandarizada y unas métricas comunes que pueden utilizarse para realizar evaluaciones de funcionamiento de módulos de detección de voz sintética.

La organización del desafío ASVspoof2015 utiliza para la evaluación una base de datos de habla natural y artificial basada en el corpus *Spoofing and Anti-Spoofing*, SAS (Wu et al., 2015b), dividida en tres juegos: entrenamiento, desarrollo y evaluación.

La voz natural contenida en la base de datos fue grabada de 106 locutores, 61 hombres y 45 mujeres, sin efectos de canal o ruidos de fondo dignos de mención. Tomando como base esas voces humanas, las voces imitadas se generan usando diferentes algoritmos de spoofing. Se utilizan para ello 10 algoritmos de spoofing diferentes, denominados S1 a S10. Entre ellos, S3 y S4 han sido creados utilizando voz sintética adaptada, uno (S10) mediante tecnología de selección de unidades, y el resto (S1, S2 y S5 a S9) conversión de voz. La mayoría de los algoritmos de síntesis adaptada y conversión de voz utilizados hacen uso del algoritmo STRAIGHT, siendo la excepción S5 que utiliza MLSA. El sistema S10, por estar basado en concatenación de unidades, no hace uso de ningún vocoder. La descripción detallada puede encontrarse en (Wu et al., 2015b) y (Wu et al., 2015c).

El material de la base de datos se divide en tren subconjuntos diferentes: entrenamiento, desarrollo y valuación, como se muestra en la Tabla 9. La división entre los diferentes subconjuntos se realiza en base a ubicar diferentes locutores en cada una de las subdivisiones.

	Número de locutores		Número de frases	
Juego	Hombres	Mujeres	Reales	Artificiales
Entrenamiento	10	15	3750	12625
Desarrollo	15	20	3497	49875
Evaluación	20	26	9404	184000

Tabla 9: Número de locutores y de frases en los distintos juegos de señales de la base de datos (Wu et al., 2015c)

6.2.1. Subconjunto de entrenamiento

Para el juego de señales de entrenamiento se seleccionaron 25 locutores, 15 mujeres y 10 hombres. Junto a las voces originales de estos locutores, forman parte de este juego sus versiones sintéticas, obtenidas mediante 5 sistemas diferentes. La cantidad de señales se resume en la Tabla 10, junto con los métodos de spoofing que incluye dicha base de datos, que son:

- Dos implementaciones de conversión de voz obtenidas utilizando STRAIGHT (Zen et al., 2007), denominados S1 y S2.
- Conversión de voz mediante MLSA (Yoshimura et al., 1999), que se referencia como S5.
- Dos implementaciones de síntesis de voces adaptadas, de nuevo utilizando STRAIGHT. Se designan como S3 y S4.

6.2.2. Subconjunto de desarrollo

El segundo subconjunto de la base de datos, diseñado para ser utilizado para el desarrollo del sistema, toma 3497 frases naturales de 35 locutores (20 mujeres y 15 hombres) y 49875 señales artificiales, generadas a partir de las naturales utilizando los mismos 5 algoritmos que se usaban para el subconjunto de entrenamiento.

6.2.3. Subconjunto de evaluación

En el subconjunto de evaluación de la base de datos se incluyen señales legítimas y falsificadas, sumando aproximadamente 200000 señales, en las mismas condiciones

que las de los juegos anteriores. Concretamente, está formado por 9404 señales humanas y 184000 imitaciones, repartidas como se detalla en la Tabla 10.

La mitad de señales simuladas fueron creadas utilizando los mismos cinco métodos denominados 'conocidos', es decir, los que ya habían sido usados en los conjuntos de señales de entrenamiento (conversión de voz mediante STRAIGHT y MLSA, y síntesis de voz adaptada usando STRAIGHT). La otra mitad de la base de datos de avaluación, sin embargo, fue generada utilizando cinco nuevas técnicas de ataque, desconocidas para los modelos creados: cuatro sistemas de conversión de voz basados en STRAIGHT y un nuevo sintetizador de voz, MaryTTS (Schröder y Trouvain, 2003), que no hace uso de vocoders. Éstos se incluyen para medir la capacidad de los sistemas SSD para enfrentarse a ataques no vistos previamente.

Subconjunto	Características	Conocido	Cantidad de señales
N	Natural	Sí	9404
S1	VC STRAIGHT	Sí	18400
S2	VC STRAIGHT	Sí	18400
S3	SS STRAIGHT	Sí	18400
S4	SS STRAIGHT	Sí	18400
S5	VC MLSA	Sí	18400
S6	VC STRAIGHT	No	18400
S7	VC STRAIGHT	No	18400
S8	VC STRAIGHT	No	18400
S9	VC STRAIGHT	No	18400
S10	Sin vocoder	No	18400

Tabla 10: Cantidad de señales utilizadas para evaluar los modelos.

6.3. Estrategias para mejorar los modelos multivocoder con información de la base de datos de ataques reales.

Al disponer de una base de datos formada por ataques concretos basados en diferentes métodos de spoofing, se pueden plantear dos grandes aproximaciones:

- Utilizar para la creación y preparación del sistema exclusivamente la base de datos de ataques realistas, completando la evaluación con la base de datos correspondiente de la misma.
- Utilizar para la creación y preparación del sistema, además de las señales detalladas en el punto anterior, información proveniente de otras bases de datos, como los modelos multivocoder que se han utilizado en secciones anteriores. Para la evaluación se utilizará el mismo juego de señales que en el punto anterior.

6.3.1. Modelado

Para la extracción de los parámetros DCT-mel-RPS que se utilizarán para el modelado y la evaluación las señales entregadas se remuestrean a 8kHz.

Para la evaluación de distintas estrategias se diseñan y prueban 4 juegos de modelos diferentes, utilizando tanto información de trabajos anteriores, como las bases de datos entregadas por la organización de ASVspoof 2015. Cada juego está formado por un modelo de habla natural (λ_{human}) y uno de spoofing (λ_{synth}).

Los dos primeros juegos, a los que llamaremos M1 y M2, se han entrenado utilizando las señales humanas y sintéticas de la nueva base de datos, para crear directamente los modelos humano y sintético. Por tanto, el modelo sintético captura las características específicas de los ataques conocidos presentes en la base de datos de entrenamiento.

Para la generación de los modelos del juego M1 se usan exclusivamente las señales del subconjunto de entrenamiento de la base de datos ASVSpoof2015. Para la creación del juego de modelos M2 se utiliza la misma idea que en M1, con la salvedad de que se utiliza toda la información disponible: las bases de datos de entrenamiento y de desarrollo, estando presentes los mismos métodos que en M1.

En el caso del juego de modelos M3, se utiliza un set multivocoder basado en WSJ y creado con la técnica descrita en secciones anteriores. El modelo humano se crea utilizando 8599 señales humanas obtenidas de 283 locutores, y las señales sintéticas se obtienen mediante transcodificación de las señales humanas utilizando los vocoders MLSA, STRAIGHT y AHOCODER. De esta forma se obtienen 25797 señales sintéticas, que se utilizan para crear el modelo de habla artificial. Para este juego de modelos no se ha utilizado ninguna señal de las originalmente entregadas por la organización del desafío ASVspoof2015.

En el caso del último juego de modelos, denominado M4, se mezclan dos estrategias diferentes: por un lado se utilizan las señales de la nueva base de datos para hacer la detección de ataques conocidos lo mejor posible, y por otro se utilizan las señales multivocoder de WSJ para mejorar la tarea de detección de ataques desconocidos. Por tanto, los modelos se crean utilizando el set de entrenamiento (utilizado en el juego M1 junto con las señales de WSJ del juego M3, sumando en total 12349 señales genuinas y 38422 artificiales.

La información sobre la cantidad de señales utilizada en cada modelo se resume en la Tabla 11.

	M1	M2	M3	M4
Natural	3750	7247	8599	12349
VC STRAIGHT	5050	12500	-	5050
VC MLSA	2525	12500	-	2525
SS STRAIGHT	5050	12500	-	5050
CS STRAIGHT	-	-	8599	8599
CS MLSA	-	-	8599	8599
CS AHOCODER	-	-	8599	8599
Total Spoofing	12625	62500	25797	38422

Tabla 11: Cantidad de señales utilizadas para entrenar los modelos, clasificadas por vocoder y método de ataque: Conversión de voz (VC), síntesis de voz adaptada (SS) y transcodificación (CS)

6.3.2. Resultados

A continuación se detallan los resultados obtenidos con esta estrategia, utilizando los métodos, modelos y bases de datos descritos.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
M1	0,506	0,234	0,031	0,039	0,241	0,654	0,010	0,103	0,011	43,638
M2	0,336	0,197	0,027	0,046	0,164	0,570	0,011	0,082	0,011	43,919
M3	19,624	17,889	0,195	0,170	11,346	13,551	7,891	0,171	16,241	49,000
M4	4,654	3,703	0,073	0,065	1,718	3,511	1,204	0,104	3,067	48,569

Tabla 12: Valores de EER en tanto por ciento de los cuatro sistemas, para cada uno de los ataques del juego de señales de evaluación.

Modelo	Resultado ante ataques conocidos	Resultado ante ataques desconocidos	Resultado ante todos los ataques
M1	0,210%	8,883%	4,547%
M2	0,154%	8,918%	4,536%
M3	9,845%	17,371%	13,608%
M4	2,042%	11,291%	6,667%

Tabla 13: Resumen agrupado de los valores de EER en tanto por ciento de los cuatro sistemas.

Tal como muestran las cifras de la Tabla 13, el uso de los modelos M1 y M2 condujo a resultados razonablemente buenos, con valores de Equal Error Rate por debajo del 0,25%, cuando la tarea se basa en detectar ataques que están presentes en los modelos. Pero cuando se trata de detectar ataques desconocidos, la tasa de error sube, acercándose al 9%. Los siguientes factores han de ser tenidos en cuenta a la hora de explicar esta degradación del funcionamiento:

- Dado que diferentes tipos de ataques pueden presentar características RPS muy diferentes, el modelar ataques específicos no asegura que los que no han sido modelados puedan ser cubiertos. De hecho, con las señales de MaryTTS (columna S10 de la Tabla 12) se obtienen unas tasas de error excepcionalmente

altas que lastran la media del sistema. Hay que tener en cuenta que MaryTTS es un sistema de síntesis basado en selección de unidades en el que no se utiliza ningún vocoder, y este tipo de sistemas quedan muy lejos del ámbito del material de entrenamiento.

- El clasificador GMM utilizado es básico, y no se ha implementado en él ningún mecanismo adicional específico para detectar ataques desconocidos.

El caso del sistema que utiliza el juego de modelos M3 requiere un análisis diferente, al haber sido creado utilizando una base de datos completamente diferente, que no incluía señales de los juegos de entrenamiento o desarrollo de la base de datos entregada por la organización. Esto provocaba que todas las señales que se utilizaban para el test fueran desconocidas, incluso las que se tomaban como ataques conocidos. El rendimiento de este sistema es pobre, con tasas de EER que llegan a superar el 10%. Algunas interpretaciones interesantes son:

- La capacidad de generalización que con estos mismos modelos se mostraba en capítulos anteriores no está funcionando con las señales de evaluación de la base de datos ASVSpoof2015.
- Pruebas internas han revelado que la verosimilitud de las señales humanas de la nueva base de datos testeadas con el modelo humano del conjunto M3 son más bajas de lo esperado. El origen de esta divergencia parece provenir de una diferencia profunda entre las señales humanas de una y otra base de datos. Entre otros factores, las condiciones de grabación pueden haber provocado severas modificaciones en la estructura de las fases.
- La mayoría de las señales artificiales del juego de evaluación se generaron usando sistemas de conversión de voz. Aunque en el capítulo 5 ha quedado establecida la capacidad del sistema para detectar voces sintéticas basadas en vocoder, con resultados que llegan al 0% de error, hasta ahora no se habían enfrentado estos modelos multivocoder con ataques basados en conversión de voces, y esta nueva situación ha de analizarse en detalle. En todo caso, dicha capacidad queda demostrada en los casos de los sistemas S3, S4 y S8. Según se puede observar en la Tabla 12, en esos casos se obtienen valores por debajo

del 0,2% de EER que son coherentes con los obtenidos en el capítulo 5 y mucho más bajos que los valores del resto de sistemas, todos ellos con valores de EER superiores al 9%

- De nuevo, el hecho de que uno de los sistemas esté basado en MaryTTS es relevante, dado que es de esperar que unos modelos basados únicamente en vocoders tengan dificultades en detectar habla sintética generada sin hacer uso de ningún vocoder.
- Por último, en los últimos años se han desarrollado diferentes versiones de los vocoders MLSA y STRAIGHT, por lo que puede haber diferencias entre los utilizados para generar las señales con que se crearon los modelos multivocoder, y aquellos con que se formaron las señales del juego de test del desafío. En la práctica, en términos de fase pueden comportarse como vocoders diferentes, haciendo que el error se incremente.

Los resultados de detección conseguidos con el juego de modelos M4 son coherentes con un diseño que modela tanto vocoders – como en M3 – como ataques específicos – como en M1 y M2-, siendo los valores de EER intermedios entre ambos.

Los resultados obtenidos por los cuatro sistemas fueron remitidos al desafío “ASVspoof 2015: Automatic Speaker Verification Spoofing and Countermeasures Challenge” (Wu et al., 2015c), cuyos resultados se resumen a continuación.

Sistema	Resultado en ataques conocidos	Resultado en ataques desconocidos	Resultado en ataques S10	Resultado sin ataques S10	Resultado en todos los ataques
A	0,408%	2,013%	8,490%	0,402%	1,211%
B	0,008%	3,922%	19,571%	0,008%	1,965%
C	0,058%	4,998%	24,601%	0,076%	2,528%
D	0,003%	5,231%	26,142%	0,003%	2,617%
E	0,041%	5,347%	26,393%	0,060%	2,694%
F	0,358%	6,078%	28,581%	0,400%	3,218%
G	0,405%	6,247%	30,021%	0,360%	3,326%
H	0,670%	6,041%	27,204%	0,706%	3,355%
I	0,005%	7,447%	37,068%	0,021%	3,726%
J	0,025%	8,168%	40,708%	0,029%	4,097%
K	0,210%	8,883%	43,638%	0,203%	4,547%
L	0,412%	13,026%	35,890%	3,478%	6,719%
M	8,528%	20,253%	31,574%	12,482%	14,391%
N	7.874%	21,262%	43,991%	11,299%	14,568%
O	17,723%	19,929%	41,519%	16,304%	18,826%
P	21,206%	21,831%	46,102%	18,786%	21,518%

Tabla 14: Resultados (EER en porcentaje) de todos los participantes en ASVspoofing challenge2015.

En la Tabla 14 se muestran los resultados de los resultados de todos los participantes en el desafío, obtenidos con el sistema cuya estrategia coincide con la nuestro modelo M1, que había que enviar obligatoriamente. El que se describe en este capítulo ha sido designado 'K'. Aunque el valor relevante para elaborar una clasificación ha sido el resultado en todos los ataques, los realmente decisivos son los ataques desconocidos, y particularmente un tipo de ataque, el S10, que ha resultado particularmente decisivo por las altas tasas de error obtenidas por todos los participantes. De hecho, el participante A, el que ha mejorado sustancialmente el reconocimiento en el sistema S10, es el que ha obtenido la mejor puntuación en el Challenge.

En el caso del sistema que se describe en este capítulo, los resultados con S10 también han lastrado el resultado total, de forma que en una clasificación en la que no se tuvieran en cuenta los resultados de este ataque se hubiera podido superar a 4 sistemas.

6.4. Estrategias de entrenamiento de modelos usando ataques reales y copy-synthesis

En los experimentos presentados en el apartado anterior, el sistema basado en un juego de modelos creado por transcodificación de la base de datos WSJ obtuvo unos resultados más pobres de lo que cabía esperar a la luz de los resultados de capítulos anteriores. Uno de los motivos aparentes era la gran diferencia existente entre las señales humanas presentes en una y otra bases de datos.

Teniendo eso en cuenta, la estrategia que va a sugerirse en esta sección utilizará el sistema de copy-synthesis mediante vocoders, pero en esta ocasión sobre las señales humanas presentes en la base de datos ASVSpoof2015.

Para la evaluación se utilizará, además del subconjunto de evaluación de la nueva base de datos, que ya se ha utilizado en 6.3, la base de datos proveniente del Blizzard Challenge que se utilizó en el capítulo 5.

Asimismo, los resultados obtenidos mediante parametrización DCT-mel-RPS se compararon con otros sistemas basados en fase como MGD (Modified Group Delay) (Zhu y Paliwal, 2004) (Hegde et al., 2007) o no, como MFCC (Imai, 1983).

6.4.1. Parametrización MGD

De las tres parametrizaciones utilizadas, DCT-mel-RPR, MFCC y MGD, las dos primeras han sido profusamente descritas en capítulos anteriores. Por tanto, en esta sección nos centraremos en la descripción de la parametrización MGD.

Los parámetros denominados Modified Group Delay (MGD) son una representación del espectro complejo de la transformada de Fourier, que contienen información tanto de la magnitud del espectro como de su fase. Se han utilizado para reconocimiento de locutor en (Zhu y Paliwal, 2004) y (Hegde et al., 2007).

Dada una señal de voz $x(n)$, la representación compleja del espectro $X(\omega)$ se puede obtener a través de la transformada de Fourier de tiempo reducido (Short Time Fourier Transform, STFT). El espectro complejo se compone de dos partes: la parte real $X_R(\omega)$ y la parte imaginaria $X_I(\omega)$. La potencia espectral de la que se derivan los populares coeficientes MFCC se representa como $|X(\omega)|^2$. Para la extracción del espectro MGD, definimos $Y(\omega)$ como el espectro complejo de $nx(n)$, una versión reescalada de la señal $x(n)$. Se define el espectro MGD $\tau_{\rho,\gamma}(\omega)$ como

$$\tau_{\rho}(\omega) = \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{|S(\omega)|^{2\rho}} \tag{29}$$

$$\tau_{\rho,\gamma}(\omega) = \frac{\tau_{\rho}(\omega)}{|\tau_{\rho}(\omega)|}|\tau_{\rho}(\omega)|^{\gamma} \tag{30}$$

donde $X_R(\omega)$ y $X_I(\omega)$ son respectivamente las partes real e imaginaria de $X(\omega)$, $Y_R(\omega)$ e $Y_I(\omega)$ las partes real e imaginaria de $Y(\omega)$, $|S(\omega)|^2$ la potencia espectral de $|X(\omega)|^2$ suavizada, y ρ y γ dos variables que controlan la forma del espectro MGD. En la práctica, $|S(\omega)|^2$ se obtiene suavizando cepstralmente la potencia espectral $|X(\omega)|^2$. Esto se consigue realizando dos pasos:

- Aplicar la transformada discreta de coseno (DCT) a la potencia espectral.
- A continuación, aplicar a los primeros 30 coeficientes DCT la transformada discreta de coseno inversa (Inverse Discrete Cosine Transform, IDCT) para reconstruir el nuevo espectro suavizado.

La motivación para usar el espectro suavizado en lugar del original estriba en que se consigue un espectro MGD más estable (Hegde et al., 2007). Se representa esta magnitud de una forma similar a un espectrograma en la Figura 13.

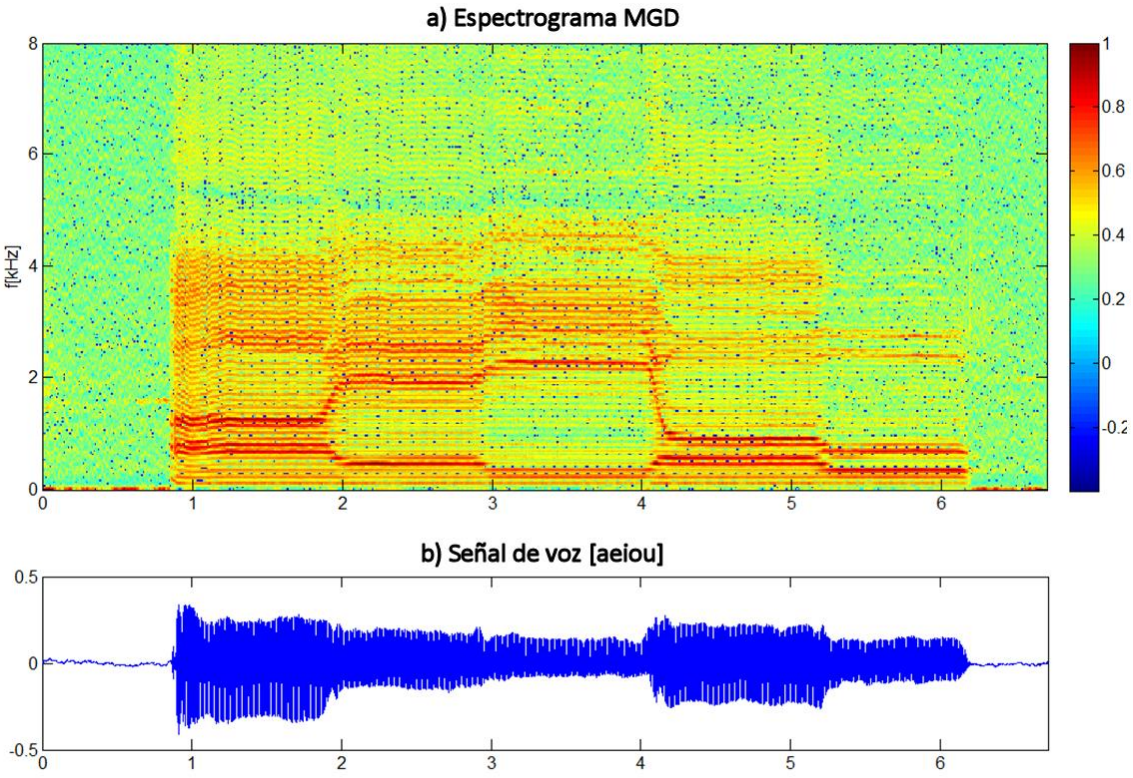

Figura 13: Espectrograma MGD de un segmento sonoro de una señal de voz, con cinco vocales consecutivas /aeiou/.

Con el espectro MGD, podemos calcular los coeficientes cepstrales MGD (Modified Group Delay Cepstral Coefficients, MGDCC) como una representación válida para el modelado. Realizaremos ese cálculo a través de los siguientes pasos:

- Aplicar la transformada de Fourier a la señal $x(n)$ y a su versión reescalada $nx(n)$ para obtener los espectro $X(\omega)$ e $Y(\omega)$, respectivamente.
- Calcular la versión suavizada $|S(\omega)|^2$ de la potencia espectral $|X(\omega)|^2$.
- Calcular el espectro MGD mediante las ecuaciones (29) y (30).
- Aplicar la DCT al espectro MGD para calcular los parámetros MGDCC

Las dos variables ρ y γ se utilizan para controlar la forma del espectro MGD. Valores pequeños de ρ incrementan la variación del espectro MGD con la frecuencia, mientras que valores pequeños de γ comprimen la amplitud del espectro MGD. Las dos variables son ajustadas manualmente para un funcionamiento optimizado, con valores de $\rho = 0{,}7$ y $\gamma = 0{,}2$.

6.4.2. Modelado

Para generar los modelos se parte de los subconjuntos de entrenamiento y desarrollo de la misma base de datos ASVSpoof2015. En lo que se refiere a los modelos que utilizará el sistema, los humanos serán siempre los mismos, mientras que se tres estrategias distintas para el entrenamiento de los modelos sintéticos, conforme a las siguientes características:

- Modelo M1: modelo sintético creado utilizando exclusivamente el material sintético de los juegos de entrenamiento y desarrollo de ASVspoof2015.
- Modelo M2: modelo sintético confeccionado utilizando señales artificiales creadas a partir de las señales humanas de los subconjuntos de entrenamiento y desarrollo de ASVspoof2015, transcodificadas mediante los vocoders AHOCODER, MLSA y STRAIGHT.
- Modelo M3: Modelo sintético creado combinando el material de los modelos M1 y M2.

En la Tabla 15 se muestra un resumen del material utilizado para entrenamiento y evaluación.

	Entrenamiento		Evaluación	
Modelo	Reales	Artificiales	Reales	Artificiales
M1	7247	62500	9404	184000
M2	7247	21741 (7247x3)	9404	184000
M3	7247	84241	9404	184000

Tabla 15: Juegos de entrenamiento y evaluación para cada uno de los 3 experimentos descritos.

6.4.3. Evaluación

Para la evaluación del rendimiento del sistema se utilizarán dos bases de datos diferentes. Por un lado, el subconjunto de evaluación de ASVspoof2015 descrito en el apartado 6.2.3, y por otro, y con el objetivo de probar el rendimiento del sistema con

señales completamente diferentes de las utilizadas para el entrenamiento, se utilizarán las señales del Blizzard Challenge (King, 2014).

Con esta base de datos se obtiene un buen número de señales basadas en diferentes sistemas de síntesis de voz punteros, ya que es el evento internacional más popular en el ámbito de evaluaciones de sistemas de síntesis de voz. Todos los participantes han de utilizar un corpus común para crear una voz sintética utilizando sus propios sistemas TTS. Para evaluar el sistema creado, se envían muestras de esa voz, que son calificadas mediante un sistema de evaluación subjetiva común en el que participa un gran número de evaluadores humanos. Por todo esto, y como ya se detalló en el apartado 5.3, se puede concluir que las señales presentes en dicho desafío son buenas representaciones de la síntesis de voz más puntera.

Adicionalmente, tras la finalización del desafío, las señales, tanto humanas como sintéticas, son puestas a disposición de la comunidad científica, de forma que son un buen campo de pruebas para evaluar un sistema SSD.

Para este experimento se han seleccionado las señales del Blizzard Challenge 2012 (King y Karaiskos, 2012). Lo conforman 11 juegos de señales, cada uno de ellos con 209 señales en inglés de Estados Unidos. Uno de los juegos está formado por señales humanas, y otro no es un participante sino un sistema estándar de selección de unidades, utilizado como referencia. Los demás, son sistemas de selección de unidades, estadísticos o híbridos

6.4.4. Resultados

Tal y como se ha descrito en el apartado 6.4.3, se realizan dos experimentos de evaluación diferentes con dos bases de datos diferentes. En ambos casos, los sistemas se han entrenado utilizando exclusivamente la información de la base de datos ASVspoof2005, concretamente los subconjuntos de entrenamiento y desarrollo, a las que se suman las señales sintéticas generadas mediante vocoders, tal como se explica en el apartado 3.2.4. En el primer experimento, el test se realiza utilizando el subconjunto de evaluación de la misma base de datos ASVspoof2015, mientras que en el segundo se utiliza una base de datos sin relación con la anterior, intentando evaluar la capacidad del sistema de enfrentarse a impostores completamente desconocidos.

6.4.4.1. Evaluación usando la base de datos ASVSpoof

En este experimento las pruebas de rendimiento del sistema SSD utilizando la parte de test de la base de datos de ASVspoof2015.

Como ya se ha mencionado, la evaluación se realizará usando señales humanas y sintéticas, generadas estas últimas mediante 10 algoritmos, 5 de ellos presentes en las señales de entrenamiento, y 5 nuevos: 4 de ellos usando conversión de voces con STRAIGHT, y el décimo, de selección de unidades, muy diferente de los entrenados.

Sistema	S1	S2	S3	S4	S5
MFCC M1	0,1102	8,4556	0,0360	0,0438	0,7621
MFCC M2	5,0327	13,1417	8,9981	8,9981	8,7375
MFCC M3	0,5471	7,6091	0,0690	0,1019	1,3040
MGD M1	0,1866	1,8559	0,2446	0,2849	2,0629
MGD M2	1,4017	4,7303	7,3400	6,9395	24,7957
MGD M3	0,4113	2,5331	0,9455	0,9500	7,4247
RPS M1	0,2661	0,1695	0,0217	0,0360	0,1439
RPS M2	1,0478	0,7359	0,1625	0,1437	0,5571
RPS M3	0,2814	0,1770	0,0152	0,0363	0,1707
Sistema	**S6**	**S7**	**S8**	**S9**	**S10**
MFCC M1	1,5449	3,4747	2,4740	0,9343	37,0711
MFCC M2	11,2048	16,3206	9,4169	9,6362	43,7628
MFCC M3	2,1734	4,3412	3,1157	1,7846	38,6373
MGD M1	2,4006	0,7891	1,3186	0,7260	39,3174
MGD M2	19,2032	2,8205	3,3713	4,3265	41,6672
MGD M3	6,5415	1,0769	1,9259	1,5708	40,1692
RPS M1	0,5147	0,0080	0,0912	0,0108	43,4680
RPS M2	2,0239	0,1286	0,2834	0,3243	47,7249
RPS M3	0,5711	0,0108	0,0755	0,0101	44,9628

Tabla 16: Valores de EER en tanto por ciento de todos los sistemas propuestos, para cada uno de los ataques del juego de señales de evaluación de ASVspoof2015.

Sistema	Ataques conocidos	Ataques conocidos TTS	Ataque conocidos VC	Ataques desconocidos excluyendo S10 (VC)
MFCC M1	1,8815	0,0399	3,1093	2,1070
MFCC M2	8,9816	8,9981	8,9707	11,6447
MFCC M3	1,9262	0,0855	3,1534	2,8537
MGD M1	0,9270	0,2647	1,3685	1,3086
MGD M2	9,0414	7,1397	10,3092	7,4304
MGD M3	2,4529	0,9478	3,4564	2,7788
RPS M1	0,1274	0,0288	0,1931	0,1562
RPS M2	0,5294	0,1531	0,7803	0,6901
RPS M3	0,1361	0,02574	0,2097	0,1669

Sistema	Ataques desconocidos	Todos los ataques excluyendo S10	Todos los ataques excluyendo S10 (σ)	Todos los ataques
MFCC M1	9,0998	1,9817	2,6973	5,4907
MFCC M2	18,0683	10,1652	3,1578	13,5250
MFCC M3	10,0104	2,3384	2,4280	5,9683
MGD M1	8,9103	1,0966	0,8445	4,9187
MGD M2	14,2777	8,3254	8,0947	11,6596
MGD M3	10,2569	2,5977	2,5715	6,3549
RPS M1	8,8185	0,1402	0,1652	4,4730
RPS M2	10,0970	0,6008	0,6171	5,3132
RPS M3	9,1261	0,1498	0,1842	4,6311

Tabla 17: Resumen de resultados (EER en porcentaje) de todos los tipos de sistemas propuestos, evaluados con la base de datos ASVspoof2015.

Los resultados de este experimento de muestran detallados para cada sistema en la Tabla 16 y agrupados por tipos de sistemas en la Tabla 17. En el anexo se muestran las curvas DET de los experimentos de la Tabla 16 realizados con RPS.

Dado que los resultados del EER medio están mediatizados por el funcionamiento degradado del sistema basado en selección de unidades S10, se presentan también resultados promediando exclusivamente el resto de los sistemas (columna "Todos los ataques excluyendo S10" en la Tabla 17). También se usará ese cálculo en el análisis que se realiza a continuación.

La referencia MFCC consigue resultados aceptables de alrededor del 2% de EER, revelando que, cuando se usa la base de datos ASVspoof2015, la tarea de clasificación no es especialmente compleja.

También es interesante señalar que –si exceptuamos el resultado de S10- no hay grandes diferencias entre ataques conocidos y desconocidos. En todos los sistemas el rendimiento cae entre el 10% y el 20% al pasar de conocido a desconocido, cuando se entrena con los propios impostores (modelo M1). Pero cuando se utiliza el modelo M2, para el que todos los ataques son desconocidos, la caída en el rendimiento es aún mayor que en los otros modelos, lo que indica que esa pequeña diferencia entre los sistemas conocidos y desconocidos no es atribuible al conocimiento 'a priori' del método de ataque, sino a circunstancias particulares de los sistemas de conversión de voz incluidos en el subconjunto de evaluación como 'desconocidos' que afectan al rendimiento. Esto se ve reforzado por los resultados de la Tabla 16, donde se puede ver que, en la mayoría de los casos, los sistemas 'desconocidos' se detectan incluso mejor que aquellos de los que se tenía conocimiento previo al entrenar los modelos. Solamente el EER del ataque S6, que tiene un error particularmente alto para cualquier modelo entrenado, independientemente de ser conocido o desconocido, hace que la puntuación media de los ataques 'desconocidos' sea más baja.

Los resultados obtenidos utilizando la parametrización MGD presentan un buen funcionamiento cuando se utilizan los modelos M1 y M3, pero la tasa de error aumenta con el modelo M2. Los parámetros MGD parecen más afectados por las

distorsiones que el proceso de modelado estadístico introduce en los algoritmos reales de spoofing, pero que no están presentes en las señales de vocoders que forman el modelo M2.

Los modelos basados en RPS, como se puede observar en la Tabla 16, consiguen unos resultados más consistentes, llegando a valores muy por debajo del 1% en todos los juegos de señales de test.

Las distintas estrategias de entrenamiento de los modelos también tienen afección sobre el rendimiento del sistema formado con parámetros RPS. El uso de los ejemplos de señales con los ataque específicos para modelar (M1) da un resultado mejor que las demás. Con el modelo M2 que utiliza exclusivamente material generado con vocoders para entrenar los modelos, el resultado es de un error algo mayor en general, pero aún razonable, en el mismo orden de magnitud de los demás. La hipotética capacidad de extrapolación de la estrategia del modelo M2 para mejorar los resultados de los sistemas desconocidos, en este caso, no ha sido tal. Tal y como se ha comentado en párrafos anteriores, la diferencia entre los sistemas designados como conocidos y desconocidos es pequeña, por lo que no podemos concluir si la capacidad de extrapolación está funcionando o no.

La estrategia M3, con un modelo sintético creado utilizando tanto con las muestras de ataques como con señales de vocoders, consigue un buen rendimiento, muy cercano al de M1. De hecho, la diferencia no es estadísticamente relevante según el test de McNemar (McNemar, 1947) (p=0,41). Sin embargo, ofrece leves mejoras en algunos de los sistemas etiquetados como desconocidos (concretamente S8 y S9).

Para comparar los niveles de error de los ataques basados en TTS con los creados mediante conversión de voz, hay que tener en cuenta que todos los ataques basados en TTS tienen representación en el subconjunto de entrenamiento, formando parte de los denominados ataques ‘conocidos’. Por tanto, la comparación válida de los sistemas TTS (columnas S3 y S4 de la Tabla 16) ha de hacerse con los sistemas ‘conocidos’ de conversión de voz. El rendimiento del sistema es claramente mejor para los ataques basados en TTS que para aquellos basados en conversión de voz. Respecto de los

ataques desconocidos (todos ellos salvo S10 basados en conversión de voces) es sorprendente comprobar, como ya se ha mencionado, que se detectan mejor que los conocidos, sobre todo en el caso de los sistemas SSD basados en MGD y MFCC.

El análisis detallado de los resultados para cada algoritmo de ataque muestra rendimientos diferentes en cada uno de los detectores. El sistema MFCC de referencia es muy sensible a las diferencias entre ataques, mostrando resultados muy dispersos (ver columna (σ) en la Tabla 17), y algunos sistemas concretos que podrían ser definidos como 'patológicos' por tener una tasa de error sobresalientemente alta (s2, S7). Para el caso de MGD, se consigue un comportamiento homogéneo con la estrategia de entrenamiento M1, pero el uso de material proveniente de vocoders (en las estrategias M2 y M3) lleva a resultados impredecibles, donde algunos de los algoritmos (S5, S6) tienen tasas de error excepcionalmente altas. En general, los sistemas SSD basados en RPS han tenido mejores resultados, y más regularidad en todos los ataques salvo S1 y S10.

6.4.4.2. Evaluación usando la base de datos de Blizzard 2012

El objetivo de este segundo experimento es analizar el rendimiento del sistema SSD cuando se enfrenta a señales totalmente desconocidas, tanto naturales como artificiales. Aparte de que el algoritmo usado para generar las señales es desconocido, estas señales también se adquirieron a través de un canal totalmente diferente, de forma que podremos evaluar la resistencia intrínseca del sistema SSD ante la problemática que genera la diversidad de canales.

Como ya se ha mencionado, utilizaremos tanto la voz natural (denominada A) como las 12 voces sintéticas adaptadas (B-K). Entre ellas, solamente 3 de los sintetizadores utilizados (concretamente E, H y K) son estadísticos y usan modelos HMM basados en ciertos parámetros. Estos, en la etapa de síntesis, alimentan un vocoder que generará la señal vocal artificial.

El resto de los sistemas se basan en selección de unidades, o tecnologías híbridas. Es decir, se concatenan segmentos de habla natural, con lo que no se utilizan vocoders. Es el mismo caso del sistema S10 correspondiente a MaryTTS de los apartados anteriores: está fuera del ámbito de los sistemas SSD que evaluamos, y requerirá un estudio

detallado en el futuro. En todo caso, la tecnología de síntesis basada en concatenación de unidades puede no ser útil para conseguir acceso a algunos sistemas de control biométrico basados en voz, por ejemplo los basados en conversación abierta (como call-centers). Para un caso así usar tecnología de selección de unidades necesitaría una base de datos de habla enorme, y sería detectable en algunos casos por las personas que lo oyeran (Wester et al., 2015).

En este experimento se evalúan todos los sistemas SSD con todas las voces disponibles y en base a tres estrategias diferentes, tal como se ve en la Tabla 18. En ella para conseguir el EER se utiliza cada uno de los sistemas TTS confrontado con el habla natural. En el anexo se muestran las curvas DET de los experimentos de la Tabla 18 realizados con RPS que obtuvieron un EER distinto de 0.

La primera conclusión evidente a la vista de las cifras es que ninguno de los sistemas seleccionados, con ninguna de las estrategias mencionadas, es capaz de detectar correctamente el habla sintética basada en selección de unidades. Lo cual es coherente con los resultados obtenidos en 6.4.4.1 por el sistema S10 basado en MaryTTS, también de selección de unidades. El error es comparable en ambos experimentos, lo que lleva a pensar que la causa de dicho error no es la falta de conocimiento previo de las señales ni otros efectos como la diferencia de canal, sino que hay una incapacidad intrínseca de detectar señales sintéticas basadas en concatenación.

En cuanto a los sistemas de síntesis basados en vocoders, los resultados dependen del sistema. El usado como baseline, basado en MFCC, consigue buenos resultados en algunos de los TTS, pero varían en función del sistema y la estrategia de entrenamiento de modelos. Para los detectores basados en MGD, el sistema es aparentemente más sensible al cambio de canal, dado que la tasa de detección no es tan buena. Por el contrario, el sistema basado en RPS, obtiene una clasificación sin errores en todos los TTS estadísticos, utilizando cualquier estrategia, lo que da una idea de la robustez frente a cambios en el canal y el método de ataque. Los errores medios para los sistemas basados en vocoder y en selección de unidades se pueden ver en la Tabla 19.

Sistema	B	C	D	E	F
MFCC M1	38,2775	44,0191	17,2249	0,0000	27,7512
MFCC M2	48,8038	48,8038	52,6316	0,9569	29,1866
MFCC M3	43,5407	45,9330	20,5742	0,0000	25,3589
MGD M1	59,8086	73,2057	8,6124	2,3923	78,4689
MGD M2	62,6794	23,4450	8,1340	5,7416	44,4976
MGD M3	59,8086	37,7990	5,2632	3,8278	60,7656
RPS M1	49,2823	69,3780	40,6699	0,0000	69,3780
RPS M2	34,9282	58,3732	11,0048	0,0000	66,0287
RPS M3	40,6699	61,7225	15,7895	0,0000	63,1579
Sistema	**G**	**H**	**I**	**J**	**K**
MFCC M1	28,7081	3,8272	22,0096	62,2010	0,0000
MFCC M2	35,4067	18,1818	24,8804	14,8325	2,8708
MFCC M3	29,1866	3,8278	22,0096	10,0478	0,0000
MGD M1	79,4258	7,6555	63,1579	42,5837	5,2632
MGD M2	32,5359	3,3493	17,7033	23,4450	3,3493
MGD M3	54,0670	3,8278	33,0144	27,2727	3,8278
RPS M1	2,8708	0,0000	32,0574	72,2488	0,0000
RPS M2	2,8708	0,0000	19,1388	6,2201	0,0000
RPS M3	3,8278	0,0000	23,9234	14,3541	0,0000

Tabla 18: Valores de EER en tanto por ciento de todos los sistemas de Blizzard confrontados a los modelos obtenidos con señales provenientes de ASVspoof2015.

Sistema	Ataque basados en vocoder (E, H, K)	Ataques híbridos y de selección de unidades	Todos los ataques
MFCC M1	1,2759	34,3131	24,4019
MFCC M2	7,3365	36,3636	27,6555
MFCC M3	1,2759	28,0930	20,0478
MGD M1	5,1037	57,8947	42,0574
MGD M2	4,1467	30,3486	22,4880
MGD M3	3,8278	39,7129	28,9474
RPS M1	0,0000	47,9836	33,5885
RPS M2	0,0000	28,3664	19,8565
RPS M3	0,0000	31,9207	22,3445

Tabla 19: Resultados (EER medio en porcentaje) para los distintos tipos de señales sintéticas.

Respecto de la estrategia de entrenamiento, el experimento muestra diferentes comportamientos según el sistema SSD utilizado y el tipo de impostor. En el sistema de referencia MFCC, la estrategia que mejores resultados ha obtenido ha sido la M3, cuyos modelos se crearon combinando señales de ataques específicos junto con señales transcodificadas mediante vocoders. Sin embargo, en los sistemas basados en MGD y RPS, es la estrategia basada en M2, que utiliza exclusivamente voz transcodificada mediante vocoders, la que obtiene mejores resultados si tenemos en cuenta el EER de todos los sistemas. Este resultado está marcado principalmente por el rendimiento obtenido en el caso de los sistemas de selección de unidades, que aunque en ninguno de los casos es suficientemente bueno, sí resulta ser algo mejor en el caso de la estrategia basada en M2. Como hipótesis se puede proponer que los modelos entrenados con ataques específicos (que conforman la estrategia M1) son demasiado específicos para conseguir capturar las características de otras técnicas de síntesis diferentes, mientras que los modelos entrenados usando señales transcodificadas por medio de vocoders, siendo de mayor calidad y más similares a las naturales, son más generales y más adecuados para la detección de señales con ataques desconocidos.

Teniendo en cuenta todos los resultados, es interesante señalar que al utilizar RPS, con información exclusivamente de fase, se consigue un resultado similar a pesar de utilizar diferentes condiciones para el entrenamiento. Sin embargo, cuando utilizamos también magnitud, como en MGD o MFCC, las diferencias entre los diferentes sistemas de entrenamiento se hacen patentes. Esto quiere decir que la magnitud del espectro queda muy distorsionada en el proceso de creación de los modelos en base a voz sintetizada o convertida, cosa que no sucede en la transcodificación mediante vocoders, en la que se obtiene una voz de alta calidad. Es, por tanto, posible, que los sistemas basados en MFCC y MGD estén modelando las mencionadas distorsiones en la magnitud del espectro. Esto explicaría los peores resultados de MFCC y MGD con la estrategia de entrenamiento M2, basada exclusivamente en señales procedentes de copy-synthesis, comparado con el resto de estrategias.

6.5. Conclusiones

En este capítulo 6 se han diseñado distintas estrategias de entrenamiento de modelos, utilizando tanto señales humanas transcodificadas como ataque específicos disponibles en la base de datos utilizada para el *ASVspoof 2015 Automatic Speaker Verification Spoofing and Countermeasures Challenge*.

En el primer diseño se desarrolla un clasificador binario basado en GMMs humanos y sintéticos creados utilizando parámetros DCT-mel-RPS en base a tres estrategias diferentes: en la primera, se utilizan las propias señales de entrenamiento entregadas en la base de datos del Challenge para entrenar los modelos humano y sintético. En el segundo, se utilizan los modelos multivocoder desarrollados para los experimentos descritos en los capítulos 4 y 5. En la tercera, se utilizan modelos combinados, usando los dos tipos de señales anteriores.

Cuando se modelan los ataques específicos, se consiguen resultados prometedores al detectar correctamente señales similares a aquellas utilizadas para crear los modelos. En el caso de los ataques desconocidos para los modelos, están profundamente lastrados por la presencia en el subconjunto de evaluación de la base de datos utilizada del sintetizador MaryTTS, basado en concatenación de unidades, y no presente en el subconjunto de entrenamiento. Pero incluso sin éste último, los

resultados de los ataques desconocidos son algo peores que los de los conocidos. Éste efecto puede explicarse porque los parámetros RPS y el clasificador GMM modelan específicamente los ataques concretos, y no son capaces de abstraerse al vocoder subyacente.

El rendimiento del sistema es modesto cuando se utilizan los modelos creados mediante uso de vocoders a partir de la base de datos WSJ para la detección de las señales que conforman la base de datos ASVspoof 2015 Challenge, debido a la gran diferencia existente entre las señales humanas de una y otra base de datos. Para mejorar estos resultados es necesario aplicar alguna técnica de adaptación de los modelos. Adicionalmente, entre los ataques presentes en el subconjunto de evaluación, está la conversión de voz, que no había sido probada con anterioridad. Y, por supuesto, la presencia del sintetizador basado en selección de unidades MaryTTS, que hace los resultados aún más pobres.

En el segundo diseño desarrollado se mantiene el planteamiento de utilizar modelos de voz transcodificada mediante vocoders, pero se utiliza como base para crearlos la voz humana de la base de datos ASVSpoof2015 en lugar de la de WSJ. Partiendo de esta base, se utilizan tres estrategias: utilizar los ataques específicos para crear los modelos, utilizar las nuevas señales codificadas, o crear modelos fusionados donde ambos tipos están presentes.

Además, la creación de los modelos se completa de tres formas distintas: mediante parámetros de fase RPS, parámetros de magnitud MFCC que se toman como referencia, y la parametrización mixta MGD. Para la evaluación se toma tanto el subconjunto de evaluación de la base de datos del Challenge, como las señales creadas para el Blizzard 2012.

Los resultados muestran que los sistemas creados de esta forma tienen un mejor rendimiento, incluso cuando éste se evalúa utilizando señales completamente desconocidas, o de otra base de datos diferente. Los dos sistemas que contienen información de fase son capaces de mejorar el sistema de referencia.

La mejor estrategia para la elaboración de modelos ha sido, a la luz de los resultados, modelar los ejemplos concretos de señales atacantes, pero también se comprueba que añadir señales transcodificadas por medio de vocoders ayuda a mejorar la detección en señales desconocidas, y en el caso concreto de los modelos basados en RPS el hecho de crear los modelos utilizando los dos tipos de señales no ha presentado inconvenientes. Sin embargo, aún se hace necesario un estudio detallado de los efectos de las distintas estrategias frente a los tipos de ataques concretos que pueden darse en un sistema real.

Los sistemas de síntesis basados en selección de unidades no son el objetivo de los detectores de voz sintética descritos en este capítulo. Sin embargo, se ha decidido mantenerlos en los tests para poder evaluar sus resultados. Tal y como cabía esperar, los detectores que usan modelos entrenados con señales generadas mediante sistemas basados en vocoders no funcionan correctamente con los ataques de sintetizadores basados en síntesis por selección de unidades. Es necesario estudiar en profundidad si los detectores basados en fase como los aquí descritos son capaces de crear modelos adecuados para detectar este tipo de amenazas en algunas circunstancias, utilizando señales adecuadas.

Los resultados de este capítulo se han presentado en (Sanchez et al., 2015a), y se han enviado también al número especial “Special Issue on Phase-Aware Signal Processing” de la revista Speech Communication, donde se encuentra en fase de revisión a la hora de escribir estas líneas.

7. Conclusiones y trabajos futuros

7.1. Aportaciones

En este trabajo se han presentado nuevas estrategias para el diseño y la implementación de técnicas para la detección de voz sintética, en el ámbito de los sistemas de verificación de locutor. Se han analizado tanto la independencia de locutor y como la independencia con el vocoder. Se ha utilizado una novedosa técnica de entrenamiento para la obtención de los modelos estadísticos del sistema para las voces artificiales utilizando copy-synthesis. Esta técnica simplifica notablemente el proceso de entrenamiento y evaluación del sistema con respecto al empleo de señales de spoofing reales. Finalmente, la validez de las estrategias y modelos propuestos se ha evaluado de forma exhaustiva utilizando ataques realistas de diferente naturaleza.

7.1.1. Validez de la parametrización RPS para SSD

Se ha diseñado un novedoso sistema SSD basado en la parametrización RPS de la fase armónica de la voz. El comportamiento del sistema se ha comparado con el uso de la parametrización de la envolvente espectral mediante MFCC, más tradicional. Comparados los resultados de ambos sistemas, el rendimiento del basado en RPS supera ampliamente al de MFCC en la mayoría de los casos, demostrando la utilidad de los parámetros de fase en la tarea de detección de voz sintética. Además, la parametrización RPS también ha funcionado mejor que otras parametrizaciones de fase tales como el MGD.

7.1.2. SSD independiente de locutor

Se ha demostrado la viabilidad de un sistema SSD basado en modelos independientes de locutor. Aunque existían algunos trabajos previos que habían experimentado con modelos independientes del locutor, estos sistemas no son separables del propio SV, sino que forman parte él y por tanto los resultados proporcionados son dependientes del sistema SV. En nuestro caso, se ha analizado por primera vez la independencia con el locutor en un sistema que realiza la tarea de detección sintética independientemente del sistema SV empleado. Esta independencia de locutor se ha validado tanto con parámetros de fase RPS como con parámetros MFCC.

7.1.3. Generación de ataques mediante transcodificación

Se ha comprobado la validez de la hipótesis de que es posible trabajar con señales transcodificadas con vocoders en lugar de con señales realistas generadas mediante TTS o conversión de voces para la obtención de los modelos de las señales sintéticas. En la mayoría de los casos estudiados la tasa de error obtenida usando modelos entrenados con señales creadas con vocoders no difiere mucho de la obtenida con los generados mediante TTS, incluso mejorando aquellas en algunos casos.

El uso de vocoders para simular los ataques de spoofing presenta importantes beneficios prácticos. Por un lado, aumenta la disponibilidad de señales al no ser necesario entrenar algoritmos de conversión de voces o de síntesis adaptada. Por otro lado, da una amplia cobertura de técnicas de suplantación, ya que muchos ataques de spoofing diferentes están basados en vocoders.

7.1.4. SSD independiente del vocoder

La dependencia de la técnica utilizada para generar las señales impostoras, representada en este caso por el vocoder utilizado, se ha analizado en profundidad. Los modelos de un solo vocoder funcionaron muy bien a la hora de detectar señales creadas con el mismo vocoder, utilizando tanto parametrización MFCC como RPS. Pero, en general, fallaban a la hora de detectar señales creadas con vocoders diferentes al del modelo. Los modelos con parametrización RPS mostraban unos resultados algo mejores a la referencia creada utilizando modelos basados en MFCC.

Para superar el problema de la dependencia del sistema con el vocoder utilizado se propone el uso de modelos multivocoder. Se ha demostrado que, utilizando parámetros RPS, aglutinar distintos vocoders en un único modelo permite mejorar la detección de señales sintéticas basadas en vocoders que no están en el modelo, con respecto al uso de un modelo basado en un único vocoder. Adicionalmente, se mantiene en límites bajos el error de detección de los vocoders que sí están en el modelo. Este efecto beneficioso de aglutinamiento de vocoders no se da con parámetros MFCC.

Tras experimentar en diferentes escenarios se llega a la conclusión de que se pueden obtener buenos resultados utilizando la parametrización RPS, mientras que mediante

los modelos basados en MFCC se llega a tasas de error mayores. Mediante la técnica de agregación de vocoders en modelos basados en parámetros RPS, se ha establecido un modelo creado con información proveniente de tres vocoders que se utiliza para enfrentarse a ataques desconocidos.

7.1.5. Validez frente a ataques realistas

El sistema SSD funcionando con los modelos multivocoder RPS ha sido capaz de detectar con éxito ejemplos reales de señales artificiales obtenidas mediante sintetizadores estadísticos desconocidos, o mediante técnicas de conversión de voces. El error de detección ha sido en la gran mayoría de los casos mucho menor al de la referencia MFCC, o al de los modelos basados en parametrización MGD. Estos buenos resultados demuestran la capacidad de extrapolación de los modelos basados en RPS y representan un indudable avance hacia un detector de voz sintética verdaderamente universal.

7.2.Trabajos futuros

Durante el desarrollo de esta tesis se han identificado líneas de investigación que pueden llevar a futuras mejoras del sistema descrito.

Por un lado, los vocoders que mantienen las fases originales de la voz, y los sistemas de síntesis que hacen uso de ellos son una amenaza real para el sistema de detección propuesto. Es necesario desarrollar acciones que puedan solucionar este problema, como podría ser entrenar el modelo multivocoder incluyendo señales generadas con este tipo de vocoders, como GlottHMM o AHOCODER-RPS.

De similar manera, el sistema es incapaz de detectar señales sintéticas generadas con sistemas TTS que no utilizan vocoders, como por ejemplo los que se basan en concatenación de forma de onda. Este tipo de señales sintéticas quedaban fuera del ámbito de este trabajo, y por tanto no se han explorado las mejoras necesarias para que un detector basado en RPS pueda tener éxito en esa tarea. Sin embargo, la importancia de esta vulnerabilidad en aplicaciones prácticas en el contexto de suplantación de voces parece baja, dado que estos sistemas de TTS no están diseñados para la adaptación de voces. Para generar una voz concreta mediante este tipo de

sintetizadores es necesario recopilar grandes bases de datos de señales de voz del locutor a imitar, lo que es muy difícil de conseguir en un escenario como el del spoofing. Pero ante la previsión de que la tecnología permita elaborar a corto plazo sintetizadores basados en concatenación de forma de onda que puedan entrenarse con una pequeña cantidad de señales del locutor a imitar, es necesario trabajar en la adaptación de los sistemas de detección para su funcionamiento con este tipo de sintetizadores.

Aunque en los capítulos 5 y 6 se han realizado experimentos en los que los modelos estaban entrenados con señales creadas en condiciones muy diferentes de las de test, consiguiendo buenos resultados de detección, no se ha realizado un estudio en profundidad del efecto del canal sobre el sistema. Hay canales de transmisión en los que se respeta la fase de la señal de voz, pero también existen otros que la distorsionan o la descartan, lo que puede afectar a la parametrización RPS. Concretamente, la transmisión telefónica puede ser un elemento distorsionador importante, tal como se estudia en (Shaw, 1997) o en (Leemann et al., 2014), y específicamente preocupante para la parametrización RPS. De igual manera ha de estudiarse el efecto de entornos ruidosos en los análisis que el clasificador SSD realiza, tal y como ya se ha realizado en los sistemas de verificación de locutor.

Por último, en este trabajo ha quedado fundamentada la capacidad de los parámetros de fase RPS para discriminar señales vocales reales y artificiales, mientras que utilizando parámetros espectrales de magnitud como MFCC puede realizarse la misma tarea aunque, como se ha descrito, se obtienen tasas de error mucho mayores. A pesar de ello, por estar basados en principios totalmente diferentes, es posible que en algunos casos concretos la parametrización MFCC pueda reconocer mejor alguna característica, siendo posible plantear el funcionamiento de un sistema fusionado que combine las capacidades de ambos para conseguir mejores resultados. Los resultados presentados en (Alam et al., 2015) para verificación de locutor y (Wang et al., 2015) para detección de voz sintética, en los que la fusión de parámetros de fase y magnitud mejora la obtenida por separado en cada uno de ellos, hace pensar que el resultado pueda ser prometedor.

7.3. Difusión de resultados

Para la difusión de los resultados de la tesis y los trabajos realizados en relación con ella se han realizado las publicaciones que se detallan a continuación.

7.3.1. Publicaciones derivadas de esta tesis

Los resultados presentados en los capítulos 3, 4 y 5 de esta tesis han sido recogidos en la siguiente publicación, que ha sido galardonada con el premio al mejor artículo 2015 de la Red Temática de Tecnologías del Habla:

- Jon Sanchez, Ibon Saratxaga, Inma Hernáez, Eva Navas, Daniel Erro y Tuomo Raitio, 2015. *Toward a Universal Synthetic Speech Spoofing Detection using Phase Information*. IEEE Transactions on Information Forensics and Security, 10, pp. 810–820.

Por otro lado, el trabajo de evaluación realizado en el capítulo 6 ha sido remitido a la revista Speech Communication bajo el título *Synthetic Speech Detection Using Phase Information*, para su publicación en un número especial con la temática específica "*Phase-Aware Signal Processing in Speech Communication*", encontrándose actualmente en la segunda fase del proceso de revisión.

7.3.2. Ponencias en congresos derivadas de esta tesis

El trabajo desarrollado en esta tesis se ha presentado en varios congresos, tanto nacionales como internacionales:

- Jon Sanchez, Ibon Saratxaga, Inma Hernáez, Eva Navas y Daniel Erro, 2014. *A Cross-vocoder Study of Speaker Independent Synthetic Speech Detection using Phase Information*. 15th Annual Conference of the International Speech Communication Association INTERSPEECH 2014. Singapur.
- Jon Sanchez, Ibon Saratxaga, Inma Hernáez, Eva Navas y Daniel Erro, 2015. *The AHOLAB RPS SSD Spoofing Challenge 2015 submission*. 16th Annual Conference of the International Speech Communication Association INTERSPEECH 2015. Dresden, Germany.

- Jon Sanchez, Ibon Saratxaga, Eva Navas, Daniel Erro e Inma Hernáez, 2015. *Fasearen erabilera ahots sintetikoaren detekzioan.* In Ikertzaile Euskaldunen Lehen Kongresua IKERGAZTE 2015. Durango.

7.3.3. Otras publicaciones

Finalmente se enumeran, ordenadas por áreas temáticas, otras publicaciones con revisión por pares en las que el autor ha participado durante el desarrollo de esta tesis:

7.3.3.1. Biometría

Publicaciones en revistas:

- J. Fierrez , J. Galbally , J. Ortega-Garcia , M. R. Freire , F. Alonso-Fernandez , D. Ramos , D. T. Toledano , J. Gonzalez-Rodriguez , J. A. Siguenza , J. Garrido-Salas , E. Anguiano-Rey , G. Gonzalez-de-Rivera , R. Ribalda , M. Faundez-Zanuy , J. A. Ortega , V. Cardeñoso-Payo , A. Viloria , C. E. Vivaracho , Q. I. Moro , J. J. Igarza , J. Sanchez , I. Hernaez , C. Orrite-Uruñuela , F. Martinez-Contreras y J. J. Gracia-Roche, 2009. *BiosecurID: A Multimodal Biometric Database*. Pattern Analysis & Applications, 12.

Otras publicaciones:

- Juan J. Igarza, Iñaki Goirizelaia, Koldo Espinosa, Inmaculada Hernáez, Raúl Méndez y Jon Sánchez, 2003. *On-line Handwritten Signature Verification Using Hidden Markov Models*. Lecture Notes in Computer Science 2095: Progress in Pattern Recognition, Speech and Image Analysis. 1, pp. 391-399.

Congresos internacionales:

- Iker Luengo, Eva Navas, Iñaki Sainz, Ibon Saratxaga, Jon Sanchez, Igor Odriozola e Inma Hernaez, 2008. *Text independent speaker identification in multilingual environments*. Sixth International Conference on Language Resources and Evaluation, LREC 2008. Marrakech, Marruecos.
- Iker Luengo, Eva Navas, Iñaki Sainz, Ibon Saratxaga, Jon Sanchez, Igor Odriozola, Juan J. Igarza e Inma Hernaez, 2008. *Building a Basque/Spanish bilingual*

database for speaker verification. Sixth International Conference on Language Resources and Evaluation, LREC 2008. Marrakech, Marruecos.

Congresos nacionales:

- David Escudero, Valentin Cardeñoso, Juan María Sanchez, Eva Navas e Inma Hernaez, 2003. *Uso de Entonación en Reconocimiento Automático de Locutor: Resultados Preliminares*. II Congreso de la Sociedad Española de Acústica Forense. Barcelona.
- Iker Luengo, Eva Navas, Inmaculada Hernáez, Jon Sanchez, Ibon Saratxaga, Iñaki Sainz y Juan J. Igarza, 2006. *Eficacia de las características prosódicas a corto plazo en la verificación de locutor*. III Jornadas en Reconocimiento Biométrico de Personas. Sevilla.

7.3.3.2. Aplicaciones de la fase armónica

Publicaciones en revistas:

- Ibon Saratxaga, Inmaculada Hernáez, Daniel Erro, Eva Navas y Jon Sánchez, 2009. *Representation of the Signal Phase for Harmonic Speech Models*. Electronic Letters, 45(7), pp. 381-383.

Otras publicaciones:

- Inma Hernaez, Ibon Saratxaga, Jianpei Ye, Jon Sanchez, Daniel Erro y Eva Navas, 2014. *Speech watermarking based on coding of the harmonic phase*. Lecture Notes in Computer Science, 8854, pp. 259-268.

Congresos internacionales:

- Inma Hernaez, Ibon Saratxaga, Jon Sanchez, Eva Navas e Iker Luengo, 2011. *Use of the Harmonic Phase in Speaker Recognition*. Interspeech 2011. Florencia, Italia.

7.3.3.3. Bases de datos orales

Publicaciones en revistas:

- Gotzon Aurrekoetxea, Karmele Fernandez-Aguirre, Jesús Rubio, Borja Ruiz y Jon Sánchez, 2013. *'DiaTech': A new tool for dialectology*. Literary and Linguistic Computing, 28, pp. 23-30.

Otras publicaciones:

- Eva Navas, Inmaculada Hernáez, Iker Luengo, Jon Sánchez e Ibon Saratxaga, 2005. *Analysis of the Suitability of Common Corpora for Emotional Speech Modeling in Standard Basque*. Lecture Notes on Computer Science, 3658, pp. 265-272.

Congresos internacionales:

- Iker Luengo, Eva Navas, Iñaki Sainz, Ibon Saratxaga, Jon Sanchez, Igor Odriozola e Inma Hernaez, 2008. *Subjective Evaluation of an Emotional Speech Database for Basque*. Sixth International Conference on Language Resources and Evaluation LREC 2008. Marrakech, Marruecos
- Gotzon Aurrekoetxea, Jon Sanchez e Igor Odriozola, 2009. *EDAK: A Corpus to Analyse Linguistic Variation*. Congreso Internacional de Lingüística del Corpus – CILC09 (Murcia 2009-05-07/09).
- Iñaki Sainz, Daniel Erro, Eva Navas, Inma Hernáez, Jon Sánchez, Ibon Saratxaga e Igor Odriozola, 2012. *Versatile Speech Databases for High Quality Synthesis for Basque*. 8th International Conference on Language Resources and Evaluation LREC 2012.
- Igor Odriozola, Eva Navas, Inma Hernáez, Iñaki Sainz, Ibon Saratxaga, Jon Sánchez and Daniel Erro, 2012. *Using an ASR database to design a pronunciation evaluation system in Basque*. 8th International Conference on Language Resources and Evaluation LREC 2012. Estambul, Turquía.

7.3.3.4. Procesado de habla en general

- Eva Navas, Inmaculada Hernáez, Amaia Castelruiz, Jon Sánchez e Iker Luengo, 2004. *Acoustical Analysis of Emotional Speech in Standard Basque for Emotions Recognition*. Lecture Notes in Computer Science 3287: Progress in Pattern Recognition, Image Analysis and Applications. 1, pp. 386-393.
- Iker Luengo, Eva Navas, Inma Hernaez y Jon Sanchez, 2005. *Reconocimiento automático de emociones utilizando parámetros prosódicos*. Revista de Procesamiento del Lenguaje Natural. 35, pp. 13-20.
- Eva Navas, Inmaculada Hernáez, Iker Luengo, Iñaki Sainz, Ibon Saratxaga y Jon Sanchez, 2007. *Meaningful Parameters in Emotion Characterisation*. Lecture Notes in Computer Science, 4775, pp. 74-84.
- Eva Navas, Iñaki Sainz, Jon Sanchez, Ibon Saratxaga e Inma Hernaez, 2009. *Algoritmo de inserción de pausas para una lengua declinada*. Revista de Procesamiento del Lenguaje Natural. 43, pp. 85-92.
- Igor Odriozola, Eva Navas, Jon Sanchez e Inma Hernaez, 2009. *Tratamiento léxico del euskara occidental basado en la división de radical y desinencia para reconocimiento de habla dialectal*. Revista de Procesamiento del Lenguaje Natural. 43, pp. 103-111.
- Iker Luengo, Eva Navas, Jon Sanchez, e Inma Hernaez, 2009. *Detección de vocales mediante modelado de clusters de fonemas*. Revista de Procesamiento del Lenguaje Natural. 43, pp. 121-128.
- Jon Sanchez, Iker Luengo, Eva Navas e Inma Hernáez, 2006. *Adaptation of the AhoTTS Text to Speech System to PDA Platfoms*. SPECOM 2006. San Petersburgo, Rusia.
- Iker Luengo, Ibon Saratxaga, Eva Navas, Inmaculada Hernáez, Jon Sanchez e Iñaki Sainz, 2007. *Evaluation of pitch detection algorithms under real conditions*. International Conference on Speech and Audio Processing (ICASSP 2007). Hawaii – EE.UU.
- Ibon Saratxaga, Inma Hernáez, Eva Navas, Iñaki Sainz, Iker Luengo, Jon Sanchez, Igor Odriozola y Daniel Erro, 2010. *AhoTransf: A Tool for Multiband Excitation*

Based Speech Analysis and Modification. Seventh international Conference on Language Resources and Evaluation LREC 2010. La Valetta, Malta.

8. Referencias

Abe, M., Nakamura, S., Shikano, K., Kuwabara, H., 1988. Voice conversion through vector quantization, en: ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing. IEEE, pp. 655-658. doi:10.1109/ICASSP.1988.196671

Aholab, 2011. AhoTTS - TTS for Basque and Spanish [WWW Document]. URL http://sourceforge.net/projects/ahotts/

Alam, J., Kenny, P., Stafylakis, T., 2015. Combining Amplitude and Phase-based Features for Speaker Verification with Short Duration Utterances, en: Interspeech 2015.

Alegre, F., Vipperla, R., Evans, N., Fauve, B., 2012. On the vulnerability of automatic speaker recognition to spoofing attacks with artificial signals. Signal Process. … 36-40.

Alku, P., 1992. Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering. Speech Commun. 11, 109-118. doi:10.1016/0167-6393(92)90005-R

Bansé, D., Doddington, G.R., Garcia-Romero, D., Godfrey, J.J., Greenberg, C.S., Martin, A.F., McCree, A., Przybocki, M., Reynolds, D.A., 2014. Summary and Initial Results of the 2013-2014 Speaker Recognition i-vector Machine Learning Challenge, en: Fifteenth Annual Conference of the International Speech Communication Association.

Beranek, B., 2013. Voice biometrics: success stories, success factors and what's next. Biometric Technol. Today 2013, 9-11. doi:10.1016/S0969-4765(13)70128-0

Beutnagel, M., Conkie, Schroeter, J., Stylianou, Y., Syrdal, A., 1999. The AT&T Next-Gen TTS System, en: Joint Meeting of ASA, EAA, and DAGA. Berlin.

Bimbot, F., Bonastre, J.F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-García, J., Petrovska-Delacrétaz, D., Reynolds, D.A., 2004. A Tutorial on Text-Independent Speaker Verification. EURASIP J. Adv. Signal Process. doi:10.1155/S1110865704310024

Black, A.W., 2006. CLUSTERGEN: A Statistical Parametric Synthesizer Using Trajectory Modeling. 9th Int. Conf. Spok. Lang. Process. 1762-1765.

Blomberg, M., Elenius, D., Zetterholm, E., 2004. Speaker verification scores and acoustic analysis of a professional impersonator. Proc. FONETIK 84-87.

Bonastre, J.F., Matrouf, D., Fredouille, C., 2007. Artificial impostor voice transformation effects on false acceptance rates, en: Interspeech.

Bredin, H., Miguel, A., Witten, I.H., Chollet, G., 2006. Detecting Replay Attacks in Audiovisual Identity Verification. 2006 IEEE Int. Conf. Acoust. Speech Signal Process. Proc. 1. doi:10.1109/ICASSP.2006.1660097

Breen, A., Jackson, P., 1998. A phonologically motivated method of selecting nonuniform units, en: Proc. ICSLP. pp. 2735-2738.

Campbell, J., Higgins, A., 1994. YOHO Speaker Verification LDC94S16. Web Download. [WWW Document]. Linguist. Data Consort. URL

https://catalog.ldc.upenn.edu/LDC94S16

Campbell, J.P., 1997. Speaker recognition: a tutorial. Proc. IEEE 85, 1437-1462. doi:10.1109/5.628714

Campbell, J.P., Shen, W., Campbell, W.M., Schwartz, R., Bonastre, J.F., Matrouf, D., 2009. Forensic speaker recognition. IEEE Signal Process. Mag. 26, 95-103.

Cappe, O., Laroche, J., Moulines, E., 1995. Regularized estimation of cepstrum envelope from discrete frequency points, en: Proceedings of 1995 Workshop on Applications of Signal Processing to Audio and Accoustics. IEEE, pp. 213-216. doi:10.1109/ASPAA.1995.482993

Chen, L., Ling, Z., Song, Y., Dai, L., 2013. Joint Spectral Distribution Modeling Using Restricted Boltzmann Machines for Voice Conversion, en: Interspeech. pp. 3052-3056.

Chen, L.-W., Guo, W., Dai, L.-R., 2010. Speaker verification against synthetic speech, en: 2010 7th International Symposium on Chinese Spoken Language Processing. IEEE, pp. 309-312. doi:10.1109/ISCSLP.2010.5684887

Coorman, G., Fackrell, J., Rutten, P., Coile, B. Van, 2000. Segment selection in the L&H Realspeak laboratory TTS system. Proc. ICSLP 395-398.

Csapo, T.G., Nemeth, G., 2014. Modeling Irregular Voice in Statistical Parametric Speech Synthesis With Residual Codebook Based Excitation. IEEE J. Sel. Top. Signal Process. 8, 209-220. doi:10.1109/JSTSP.2013.2292037

De Leon, P.L., Apsingekar, V.R., Pucher, M., 2010a. REVISITING THE SECURITY OF SPEAKER VERIFICATION SYSTEMS AGAINST IMPOSTURE USING SYNTHETIC SPEECH 1798-1801.

De Leon, P.L., Hernáez, I., Saratxaga, I., Pucher, M., Yamagishi, J., 2011. Detection of Synthetic Speech for the Problem of Imposture. 2011 IEEE Int. Conf. Acoust. Speech Signal Process. ICASSP 4844-4847.

De Leon, P.L., Pucher, M., Yamagishi, J., 2010b. Evaluation of the vulnerability of speaker verification to synthetic speech, en: Proc. Odyssey 2010, The Speaker and Language Recognition Workshop. pp. 151-158.

De Leon, P.L., Pucher, M., Yamagishi, J., Hernáez, I., Saratxaga, I., 2012. Evaluation of Speaker Verification Security and Detection of HMM-Based Synthetic Speech. IEEE Trans. Audio. Speech. Lang. Processing 20, 2280-2290. doi:10.1109/TASL.2012.2201472

Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P., 2011. Front-End Factor Analysis for Speaker Verification. IEEE Trans. Audio. Speech. Lang. Processing 19, 788-798. doi:10.1109/TASL.2010.2064307

Desai, S., Black, A.W., Yegnanarayana, B., Prahallad, K., 2010. Spectral Mapping Using Artificial Neural Networks for Voice Conversion. IEEE Trans. Audio. Speech. Lang. Processing 18, 954-964. doi:10.1109/TASL.2010.2047683

Donovan, R.E., Eide, E.M., 1998. The IBM Trainable Speech Synthesis System. 5th Int.

Conf. Spok. Lang. Process. 5, 1703-1706.

Drugman, T., Moinet, A., Dutoit, T., Wilfart, G., 2009. Using a pitch-synchronous residual codebook for hybrid HMM/frame selection speech synthesis, en: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 3793-3796. doi:10.1109/ICASSP.2009.4960453

Drygajlo, A., 2007. Forensic automatic speaker recognition. IEEE Signal Process. Mag. 24, 132-135.

Dutoit, T., Holzapfel, A., Jottrand, M., Moinet, A., Perez, J.M., Stylianou, Y., 2007. Towards a Voice Conversion System Based on Frame Selection, en: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2007). IEEE, pp. IV-513-IV-516. doi:10.1109/ICASSP.2007.366962

Elenius, K., Lindberg, J., 1997. SpeechDat Speech Databases for Creation of Voice Driven Teleservices. Phonum 4, Phonetics 61-64.

Endres, W., Bambach, W., Flösser, G., 1971. Voice spectrograms as a function of age, voice disguise, and voice imitation. J. Acoust. Soc. Am. 49, 1842-1848.

Eriksson, E.J., Sullivan, K.P.H., Zetterholm, E., Czigler, P.E., Green, J., Skagerstrand, Å., Van Doorn, J., 2010. Detection of imitated voices, who are reliable earwitnesses? Int. J. Speech Lang. Law.

Erro, D., 2008. Intra-lingual and cross-lingual voice conversion using harmonic plus stochastic models. UPC.

Erro, D., Navas, E., Hernáez, I., 2013. Parametric Voice Conversion Based on Bilinear Frequency Warping Plus Amplitude Scaling. IEEE Trans. Audio. Speech. Lang. Processing 21, 556-566. doi:10.1109/TASL.2012.2227735

Erro, D., Sainz, I., Navas, E., Hernáez, I., 2014. Harmonics Plus Noise Model Based Vocoder for Statistical Parametric Speech Synthesis. IEEE J. Sel. Top. Signal Process. 8, 184-194. doi:10.1109/JSTSP.2013.2283471

Erro, D., Sainz, I., Navas, E., Hernáez, I., 2011. Improved HNM-Based Vocoder for Statistical Synthesizers., en: Interspeech. Florence, Italy, pp. 1809 - 1812.

Evans, N., Alegre, F., Wu, Z., Kinnunen, T., 2014. Anti-spoofing: Voice conversion. Book chapter in «Encyclopedia of Biometrics», 2nd Edition, Springer, Stan Z. Li and Anil K. Jain, Eds, September 9th, 2014. doi:http://dx.doi.org/10.1007/978-3-642-27733-7_9111-2

Evans, N.W.D., Kinnunen, T., Yamagishi, J., 2013. Spoofing and countermeasures for automatic speaker verification, en: INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association. London.

Farrús, M., Wagner, M., Anguita, J., Hernando, J., 2008. How vulnerable are prosodic features to professional imitators?, en: Odyssey 2008: The Speaker Recognition Workshop. Stellenbosch, South Africa.

Farrús, M., Wagner, M., Erro, D., Hernando, J., 2010. Automatic Speaker Recognition as a Measurement of Voice Imitation and Conversion. Int. J. Speech Lang. Law 17.

doi:10.1558/ijsll.v17i1.119

Faundez-Zanuy, M., Monte-Moreno, E., 2005. State-of-the-art in speaker recognition. IEEE Aerosp. Electron. Syst. Mag. 20, 7-12. doi:10.1109/MAES.2005.1432568

Fierrez, J., Galbally, J., Ortega-Garcia, J., Freire, M.R., Alonso-Fernandez, F., Ramos, D., Toledano, D.T., Gonzalez-Rodriguez, J., Siguenza, J.A., Garrido-Salas, J., Anguiano, E., Gonzalez-de-Rivera, G., Ribalda, R., Faundez-Zanuy, M., Ortega, J.A., Cardeñoso-Payo, V., Viloria, A., Vivaracho, C.E., Moro, Q.I., Igarza, J.J., Sanchez, J., Hernáez, I., Orrite-Uruñuela, C., Martinez-Contreras, F., Gracia-Roche, J.J., 2009. BiosecurID: a multimodal biometric database. Pattern Anal. Appl. 13, 235-246. doi:10.1007/s10044-009-0151-4

Fisher, W.M., Doddington, G.R., Goudie-Marshall, K.M., 1986. The DARPA speech recognition research database: specifications and status, en: Proc. DARPA Workshop on speech recognition. pp. 93-99.

Fredouille, C., Charlet, D., 2014. Analysis of I-Vector Framework for Speaker Identification in TV-Shows, en: Fifteenth Annual Conference of the International Speech Communication Association.

Furui, S., 1981. Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust. 29, 254-272. doi:10.1109/TASSP.1981.1163530

Galou, G., Chollet, G., 2011. Synthetic Voice Forgery in the Forensic Context: a short tutorial. Forensic Speech Audio Anal. Work. Gr. 3-5.

Gillett, B., King, S., 2003. Transforming F0 Contours, en: Proc. Eurospeech 2003. International Speech Communication Association, Geneva.

Godoy, E., Rosec, O., Chonavel, T., 2012. Voice Conversion Using Dynamic Frequency Warping With Amplitude Scaling, for Parallel or Nonparallel Corpora. IEEE Trans. Audio. Speech. Lang. Processing 20, 1313-1323. doi:10.1109/TASL.2011.2177820

González Hautamäki, R., Kinnunen, T., Hautamäki, V., Leino, T., Laukkanen, A., 2013. I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry, en: INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association. Lyon, France, pp. 930-934.

Greenberg, C.S., Bansé, D., Doddington, G.R., Garcia-romero, D., Godfrey, J.J., Kinnunen, T., Martin, A.F., Mccree, A., Przybocki, M., Reynolds, D.A., 2014. The NIST 2014 Speaker Recognition i-Vector Machine Learning Challenge, en: Odyssey 2014: The Speaker and Language Recognition Workshop. pp. 224-230.

Hegde, R.M., Murthy, H.A., Gadde, V.R.R., 2007. Significance of the Modified Group Delay Feature in Speech Recognition. IEEE Trans. Audio, Speech Lang. Process. 15, 190-202. doi:10.1109/TASL.2006.876858

Helander, E.E., Nurminen, J., 2007. A Novel Method for Prosody Prediction in Voice Conversion, en: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07. IEEE, pp. IV-509-IV-512. doi:10.1109/ICASSP.2007.366961

Hernáez, I., Saratxaga, I., Sanchez, J., Navas, E., Luengo, I., 2011. Use of The Harmonic

Phase in Speaker Recognition, en: INTERSPEECH 2011, 12 th Annual Conference of the International Speech Communication Association. Florence, Italy, pp. 2757-2760.

HTS Working Group, 2002. HMM-based Speech Synthesis System (HTS) [WWW Document]. URL http://hts.sp.nitech.ac.jp/

Hunt, A.J., Black, A.W., 1996. Unit selection in a concatenative speech synthesis system using a large speech database, en: 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. IEEE, pp. 373-376. doi:10.1109/ICASSP.1996.541110

Imai, S., 1983. Cepstral analysis synthesis on the mel frequency scale, en: ICASSP '83. IEEE International Conference on Acoustics, Speech, and Signal Processing. Institute of Electrical and Electronics Engineers, pp. 93-96. doi:10.1109/ICASSP.1983.1172250

Jain, A.K., Ross, A., Pankanti, S., 2006. Biometrics: A Tool for Information Security. IEEE Trans. Inf. Forensics Secur. 1, 125-143. doi:10.1109/TIFS.2006.873653

Kain, A., Macon, M.W., 1998. Spectral voice conversion for text-to-speech synthesis, en: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181). IEEE, pp. 285-288. doi:10.1109/ICASSP.1998.674423

Kawahara, H., Masuda-Katsuse, I., de Cheveigné, A., 1999. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. Speech Commun. 27, 187-207. doi:10.1016/S0167-6393(98)00085-5

Khoury, E., Kinnunen, T., Sizov, A., Wu, Z., Marcel, S., 2014. Introducing I-Vectors for Joint Anti-spoofing and Speaker Verification, en: The 15th Annual Conference of the International Speech Communication Association.

King, S., 2014. Measuring a decade of progress in Text-to-Speech. Loquens 1, e006. doi:10.3989/loquens.2014.006

King, S., Clark, R.A.J., Mayo, C., Karaiskos, V., 2008. The Blizzard Challenge 2008.

King, S., Karaiskos, V., 2012. The Blizzard Challenge 2012, en: Proc. of The Blizzard Challenge 2012.

King, S., Karaiskos, V., 2011. The Blizzard Challege 2011, en: Proc. of The Blizzard Challenge 2011. Torino, Italy.

Kinnunen, T., Wu, Z., Lee, K.A., Sedlak, F., Chng, E.S., Li, H., 2012. Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech, en: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4401-4404. doi:10.1109/ICASSP.2012.6288895

Klatt, D.H., 1980. Software for a cascade/parallel formant synthesizer. J. Acoust. Soc. Am. 67, 971. doi:10.1121/1.383940

Kurematsu, A., Takeda, K., Sagisaka, Y., Katagiri, S., Kuwabara, H., Shikano, K., 1990. ATR Japanese speech database as a tool of speech recognition and synthesis. Speech Commun. 9, 357-363. doi:10.1016/0167-6393(90)90011-W

Larcher, A., Lee, K., Ma, B., Li, H., 2012. RSR2015: Database for Text-Dependent Speaker Verification using Multiple Pass-Phrases. INTERSPEECH 2-5.

Lau, Y.W., Tran, D., Wagner, M., 2005. Testing Voice Mimicry with the YOHO Speaker Verification Corpus, en: Proceedings of the 9th International Conference on Knowledge-Based Intelligent Information and Engineering Systems - Volume Part IV, KES'05. Springer-Verlag, Berlin, Heidelberg, pp. 15-21. doi:10.1007/11554028_3

Lau, Y.W., Wagner, M., Tran, D., 2004. Vulnerability of speaker verification to voice mimicking. Proc. 2004 Int. Symp. Intell. Multimedia, Video Speech Process. 2004. doi:10.1109/ISIMP.2004.1434021

Leemann, A., Kolly, M.-J., Dellwo, V., 2014. Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison. Forensic Sci. Int. 238, 59-67. doi:10.1016/j.forsciint.2014.02.019

Ling, Z., Wu, Y., Wang, Y., Qin, L., Wang, R., 2006. USTC System for Blizzard Challenge 2006 an Improved HMM-based Speech Synthesis Method. Blizzard Chall. Work. 4-7.

Lolive, D., Barbot, N., Boëffard, O., 2008. Pitch and Duration Transformation with Non-parallel Data, en: Proceedings of the Speech Prosody 2008 Conference. pp. 111-114.

Luengo, I., 2010. Análisis y Evaluación de Parámetros para Identificación Automática de Emociones en el Habla.

Luengo, I., Saratxaga, I., Navas, E., Hernaez, I., Sanchez, J., Sainz, I., 2007. Evaluation of Pitch Detection Algorithms Under Real Conditions, en: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07. IEEE, pp. IV-1057-IV-1060. doi:10.1109/ICASSP.2007.367255

Maia, R., Akamine, M., Gales, M.J.F., 2012. Complex cepstrum as phase information in statistical parametric speech synthesis, en: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4581-4584. doi:10.1109/ICASSP.2012.6288938

Maia, R., Toda, T., Zen, H., Nankaku, Y., Tokuda, K., 2007. An excitation model for HMM-based speech synthesis based on residual modeling. Procs. 6th ISCA Work. Speech Synth. 131-136.

Mariéthoz, J., Bengio, S., 2005. Can a Professional Imitator Fool a GMM-Based Speaker Verification System?

Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., 1997. The DET Curve in Assessment of Detection Task Performance.

Masuko, T., Hitotsumatsu, T., Tokuda, K., Kobayashi, T., 1999. On the security of HMM-based speaker verification systems against imposture using synthetic speech, en:

Proceedings of the European Conference on Speech Communication and Technology. Citeseer, pp. 1223–1226. doi:10.1.1.61.2528

Masuko, T., Tokuda, K., Kobayashi, T., 2000. Imposture Using Synthetic Speech Against Speaker Verification Based On Spectrum And Pitch, en: Proc. Int. Conf. Spoken Lang. Process. (ICSLP). pp. 302-605.

Matrouf, D., Bonastre, J.F., Fredouille, C., 2006. Effect of Speech Transformation on Impostor Acceptance, en: 2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings. IEEE, pp. I-933-I-936. doi:10.1109/ICASSP.2006.1660175

McNemar, Q., 1947. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12, 153-157.

Melin, H., 1996. Gandalf-a Swedish telephone speaker verification database. Proceeding Fourth Int. Conf. Spok. Lang. Process. ICSLP '96 3. doi:10.1109/ICSLP.1996.608018

Miguel, A., Villalba, J., Ortega, A., Lleida, E., Vaquero, C., 2014. Factor Analysis with Sampling Methods for Text Dependent Speaker Recognition. Proc. 15th Annu. Conf. Int. Speech Commun. Assoc. Interspeech 2014 1342-1346.

Monte-Moreno, E., Chetouani, M., Faundez-Zanuy, M., Sole-Casals, J., 2009. Maximum likelihood linear programming data fusion for speaker recognition. Speech Commun. 51, 820-830. doi:10.1016/j.specom.2008.05.009

Moulines, E., Charpentier, F., 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Commun. 9, 453-467. doi:10.1016/0167-6393(90)90021-Z

Neiberg, D., 2001. Text Independent speaker verification using adapted Gaussian mixture models. Cent. Speech Technol. Dep. Speech, Music Hear. KTH, Stock. Sweden 11-12.

NIST, 2010. The NIST Year 2010 Speaker Recognition Evaluation Plan [WWW Document]. URL www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf

NIST, 2008. The NIST Year 2008 Speaker Recognition Evaluation Plan [WWW Document]. URL http://www.itl.nist.gov/iad/mig//tests/sre/2008/sre08_evalplan_release4.pdf

NIST, 2007. NIST Evaluation Tools [WWW Document]. URL http://www.itl.nist.gov/iad/mig//tools/

NIST, 2006. The NIST Year 2006 Speaker Recognition Evaluation Plan [WWW Document]. URL http://www.itl.nist.gov/iad/mig//tests/sre/2006/sre-06_evalplan-v9.pdf

NIST, 2005. The NIST Year 2005 Speaker Recognition Evaluation Plan [WWW Document]. URL http://www.itl.nist.gov/iad/mig//tests/sre/2005/sre-05_evalplan-v6.pdf

NIST, 2004. The NIST Year 2004 Speaker Recognition Evaluation Plan [WWW Document]. URL http://www.itl.nist.gov/iad/mig//tests/sre/2004/SRE-04_evalplan-v1a.pdf

NIST, 2002. The NIST Year 2002 Speaker Recognition Evaluation Plan [WWW Document]. URL http://www.itl.nist.gov/iad/mig//tests/sre/2002/2002-spkrec-evalplan-v60.pdf

NIST, 1995. NIST Speaker Recognition Evaluation [WWW Document]. URL http://www.itl.nist.gov/iad/mig//tests/sre/

Novoselov, S., Pekhovsky, T., Simonchik, K., Shulipa, A., 2014. RBM-PLDA Subsystem for the NIST i-Vector Challenge, en: Fifteenth Annual Conference of the International Speech Communication Association.

Ogihara, A., Unno, H., Shiozaki, A., 2005. Discrimination Method of Synthetic Speech Using Pitch Frequency against Synthetic Speech Falsification. EICE Trans. Fundam. Electron. Commun. Comput. Sci. E88-A, 280-286.

Ohm, G.S., 1843. Ueber die Definition des Tones, nebst daran geknüpfter Theorie der Sirene und ähnlicher tonbildender Vorrichtungen. Ann. Phys. 135, 513-565. doi:10.1002/andp.18431350802

Ortega-Garcia, J., Gonzalez-Rodriguez, J., Marrero-Aguiar, V., 2000. AHUMADA: A large speech corpus in Spanish for speaker characterization and identification. Speech Commun. 31, 255-264.

Paalanen, P., Kämäräinen, J., Ilonen, J., Kälviäinen, H., 2006. Feature Representation and Discrimination Based on Gaussian Mixture Model Probability Densities – Practices and Algorithms. Pattern Recognit. 39, 1346-1358. doi:10.1016/j.patcog.2006.01.005

Paul, D.B., Baker, J.M., 1992. The design for the wall street journal-based CSR corpus, en: Proceedings of the workshop on Speech and Natural Language - HLT '91. Association for Computational Linguistics, Morristown, NJ, USA, p. 357. doi:10.3115/1075527.1075614

Pellom, B.L., Hansen, J.H.L., 1999. An experimental study of speaker verification sensitivity to computer voice-altered imposters, en: 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258). pp. 837-840. doi:10.1109/ICASSP.1999.759801

Perrot, P., Aversano, G., Blouet, R., Charbit, M., Chollet, G., 2005. Voice Forgery Using ALISP: Indexation in a Client Memory., en: ICASSP (1). pp. 17-20.

Pinto, R.G.C.P., Pinto, H.L.C.P., Caloba, L.P., 1995. Using neural networks for automatic speaker recognition: a practical approach. 38th Midwest Symp. Circuits Syst. Proc. 2.

Qian, Y., Fan, Y., Hu, W., Soong, F.K., 2014. On the training aspects of Deep Neural Network (DNN) for parametric TTS synthesis, en: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 3829-3833. doi:10.1109/ICASSP.2014.6854318

Quatieri, T.F., 2002. Discrete-time speech signal processing: principles and practice. Pearson Education India.

Raitio, T., Suni, A., Pulakka, H., Vainio, M., Alku, P., 2011a. Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis, en: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4564-4567. doi:10.1109/ICASSP.2011.5947370

Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M., Alku, P., 2011b. HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering. IEEE Trans. Audio. Speech. Lang. Processing 19, 153-165. doi:10.1109/TASL.2010.2045239

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker Verification Using Adapted Gaussian Mixture Models. Digit. Signal Process. 10, 19-41. doi:10.1006/dspr.1999.0361

Rosenberg, A.E., 1976. Automatic speaker verification: A review. Proc. IEEE 64, 475-487. doi:10.1109/PROC.1976.10156

Sanchez, J., Saratxaga, I., Hernáez, I., Navas, E., Erro, D., 2015a. The AHOLAB RPS SSD Spoofing Challenge 2015 submission, en: INTERSPEECH 2015, 16 th Annual Conference of the International Speech Communication Associationth Annual Conference of the International Speech Communication Association. Dresden, Germany.

Sanchez, J., Saratxaga, I., Hernáez, I., Navas, E., Erro, D., 2014. A Cross-vocoder Study of Speaker Independent Synthetic Speech Detection using Phase Information, en: Interspeech. Singapore, pp. 1663-1667.

Sanchez, J., Saratxaga, I., Hernáez, I., Navas, E., Erro, D., Raitio, T., 2015b. Toward a Universal Synthetic Speech Spoofing Detection using Phase Information. IEEE Trans. Inf. Forensics Secur. PP, 1-1. doi:10.1109/TIFS.2015.2398812

Saratxaga, I., 2011. La fase en los modelos armónicos de la señal de voz: estrategias de representación, tratamiento y aplicaciones. Euskal Herriko Unibertsitatea - Universidad del Pais Vasco.

Saratxaga, I., Erro, D., Hernáez, I., Sainz, I., Navas, E., 2009a. Use of Harmonic Phase Information for Polarity Detection in Speech Signals, en: Interspeech. ISCA, pp. 1075-1078.

Saratxaga, I., Hernáez, I., Erro, D., Navas, E., Sanchez, J., 2009b. Simple representation of signal phase for harmonic speech models. Electron. Lett. 45, 381. doi:10.1049/el.2009.3328

Saratxaga, I., Hernáez, I., Odriozola, I., Navas, E., Luengo, I., Erro, D., 2010. Using harmonic phase information to improve ASR rate., en: Proc. Interspeech 2010. Makuhari, Japan, pp. 1185 - 1188.

Saratxaga, I., Hernáez, I., Pucher, M., Navas, E., Sainz, I., 2012. Perceptual Importance of the Phase Related Information in Speech, en: Interspeech. ISCA, Portland, OR, pp. 1448-1451. doi:10.1.1.396.4789

Satoh, T., Masuko, T., Kobayashi, T., Tokuda, K., 2001. A robust speaker verification

system against imposture using an HMM-based speech synthesis system., en: Dalsgaard, P., Lindberg, B., Benner, H., Tan, Z.-H. (Eds.), Interspeech. ISCA, pp. 759-762.

Schröder, M., Trouvain, J., 2003. The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching, en: International Journal of Speech Technology. pp. 365-377.

Shang, W., Stevenson, M., 2008. A playback attack detector for speaker verification systems, en: 2008 3rd International Symposium on Communications, Control, and Signal Processing, ISCCSP 2008. pp. 1144-1149.

Shaw, A., 1997. Voice verification — Authenticating remote users over the telephone. Netw. Secur. 1997, 16-18. doi:10.1016/S1353-4858(97)83241-2

Soong, F., Rosenberg, A., Rabiner, L., Juang, B., 1985. A vector quantization approach to speaker recognition. ICASSP '85. IEEE Int. Conf. Acoust. Speech, Signal Process. 10.

Sorin, A., Shechtman, S., Pollet, V., 2011. Uniform speech parameterization for multi-form segment synthesis. … Annu. Conf. Int. Speech … 337-340.

SPTK, 2014. Speech Signal Processing Toolkit (SPTK) [WWW Document]. URL http://sp-tk.sourceforge.net/

Stevens, S.S., Volkmann, J., Newman, E.B., 1937. The mel scale equates the magnitude of perceived differences in pitch at different frequencies. J. Acoust. Soc. Am 8, 185-190.

Steward, B., De Leon, P.L., Yamagishi, J., 2012. Synthetic speech discrimination using pitch pattern statistics derived from image analysis, en: Interspeech. pp. 370-373.

Stylianou, Y., 2009. Voice Transformation: A survey, en: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 3585-3588. doi:10.1109/ICASSP.2009.4960401

Stylianou, Y., 1996. Harmonic plus Noise models for Speech, combined with Statistical Methods, for Speech and Speaker Modification. PhD Thesis Ec. Natl. Supèrieure des Télécommunications Paris. doi:citeulike-article-id:6662479

Stylianou, Y., Cappe, O., Moulines, E., 1998. Continuous probabilistic transform for voice conversion. IEEE Trans. Speech Audio Process. 6, 131-142. doi:10.1109/89.661472

Sündermann, D., Höge, H., Bonafonte, A., Ney, H., Black, A., Narayanan, S., 2006. Text-independent voice conversion based on unit selection, en: Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on. pp. I-I.

Sundermann, D., Ney, H., 2003. VTLN-based voice conversion. IEEE, pp. 556-559. doi:10.1109/ISSPIT.2003.1341181

Taylor, P.A., Black, A., Caley, R., 1998. The Architecture of the Festival Speech Synthesis System, en: The Third ESCA Workshop in Speech Synthesis. Jenolan Caves,

Australia, pp. 147-151.

Toda, T., Black, A.W., Tokuda, K., 2007. Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory. IEEE Trans. Audio, Speech Lang. Process. 15, 2222-2235. doi:10.1109/TASL.2007.907344

Toda, T., Saruwatari, H., Shikano, K., 2001. Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum, en: 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221). IEEE, pp. 841-844. doi:10.1109/ICASSP.2001.941046

Tokuda, K., Kobayashi, T., Masuko, T., Imai, S., 1994. Mel-Generalized Cepstral Analysis - a Unified Approach To Speech Spectral Stimation, en: Proc. Int. Conf. Spoken Lang. Process. (ICSLP). pp. 1043-1046.

Villalba, J., Lleida, E., 2011a. Detecting replay attacks from far-field recordings on speaker verification systems, en: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 274-285.

Villalba, J., Lleida, E., 2011b. Preventing replay attacks on speaker verification systems, en: Proceedings - International Carnahan Conference on Security Technology.

Villalba, J., Lleida, E., Ortega, A., Miguel, A., 2013. A new bayesian network to assess the reliability of speaker verification decisions. Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH 3132-3136.

Wang, L., Yoshida, Y., Kawakami, Y., Nakagawa, S., 2015. Relative phase information for detecting human speech and spoofed speech, en: Interspeech 2015. Dresden, Germany.

Wester, M., Wu, Z., Yamagishi, J., 2015. Human vs Machine Spoofing Detection on Wideband and Narrowband Data, en: Interspeech 2015. Singapore, pp. 1-5.

Wu, C.-H., Hsia, C.-C., Liu, T.-H., Wang, J.-F., 2006. Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis. IEEE Trans. Audio, Speech Lang. Process. 14, 1109-1116. doi:10.1109/TASL.2006.876112

Wu, Z., 2015. SASCorpus [WWW Document]. URL https://wiki.inf.ed.ac.uk/CSTR/SASCorpus (accedido 2.25.15).

Wu, Z., Chng, E.S., Li, H., 2012. Detecting Converted Speech and Natural Speech for anti-Spoofing Attack in Speaker Recognition. Interspeech 2-5.

Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., Li, H., 2015a. Spoofing and countermeasures for speaker verification: A survey. Speech Commun. 66, 130-153. doi:10.1016/j.specom.2014.10.005

Wu, Z., Khodabakhsh, A., Demiroglu, C., Yamagishi, J., King, S., 2015b. SAS : A SPEAKER VERIFICATION SPOOFING DATABASE CONTAINING DIVERSE ATTACKS, en: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP) 2015.

Wu, Z., Kinnunen, T., Evans, N., Yamagishi, J., Hanilci, C., Sahidullah, M., Sizov, A.,

2015c. ASVspoof 2015: the First Automatic Speaker Verification Spoofing and Countermeasures Challenge, en: Poc. Interspeech 2015.

Wu, Z., Xiao, X., Chng, E.S., Li, H., 2013. Synthetic speech detection using temporal modulation feature, en: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 7234-7238. doi:10.1109/ICASSP.2013.6639067

Yamagishi, J., Nose, T., Zen, H., Ling, Z.-H., Toda, T., Tokuda, K., King, S., Renals, S., 2009. Robust Speaker-Adaptive HMM-Based Text-to-Speech Synthesis. IEEE Trans. Audio. Speech. Lang. Processing 17, 1208-1230. doi:10.1109/TASL.2009.2016394

Yoshimura, T., Tokuda, K., Kobayashi, T., Masuko, T., Kitamura, T., 1999. Simultaneous Modeling Of Spectrum, Pitch And Duration In HMM-Based Speech Synthesis, en: Eurospeech. pp. 2347-2350.

Zen, H., Toda, T., Nakamura, N., Tokuda, K., 2007. Details of the Nitech HMM-Based Speech Synthesis System for the Blizzard Challenge 2005. IEICE Trans. Inf. Syst. E90-D, 325-333. doi:10.1093/ietisy/e90-1.1.325

Zetterholm, E., 2007. Detection of Speaker Characteristics Using Voice Imitation. Speak. Classif. II 4441, 192-205. doi:10.1007/978-3-540-74122-0_16

Zhu, D., Paliwal, K.K., 2004. Product of power spectrum and group delay function for speech recognition, en: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, pp. I-125-8. doi:10.1109/ICASSP.2004.1325938

9. Anexo: Curvas DET

Se muestran a continuación las curvas DET obtenidas en los experimentos de evaluación del capítulo 6, destinados a establecer las mejores estrategias de entrenamiento de los modelos.

Las curvas DET se han obtenido mediante el software DETware de (NIST, 2007).

9.1. Estrategias elaboradas usando ataques específicos y copy-synthesis sobre la base de datos WSJ

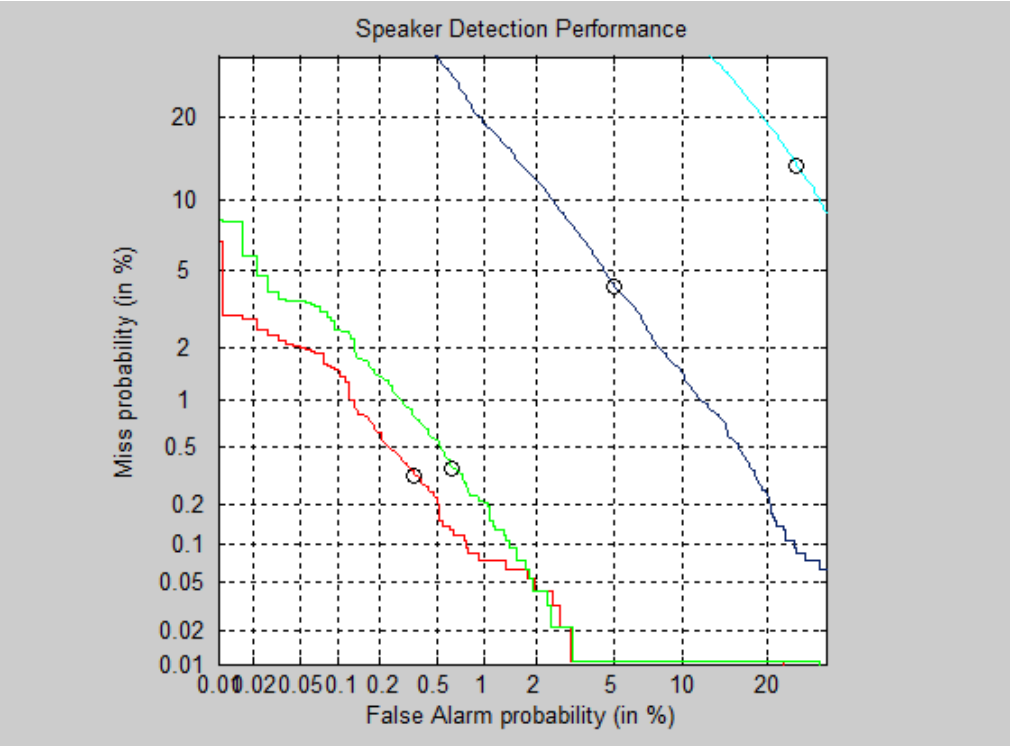

Figura 14: Curva DET del experimento de detección del sistema S1 usando los 4 juegos de modelos RPS: M1 en verde, M2 en rojo, M3 en azul y M4 en gris.

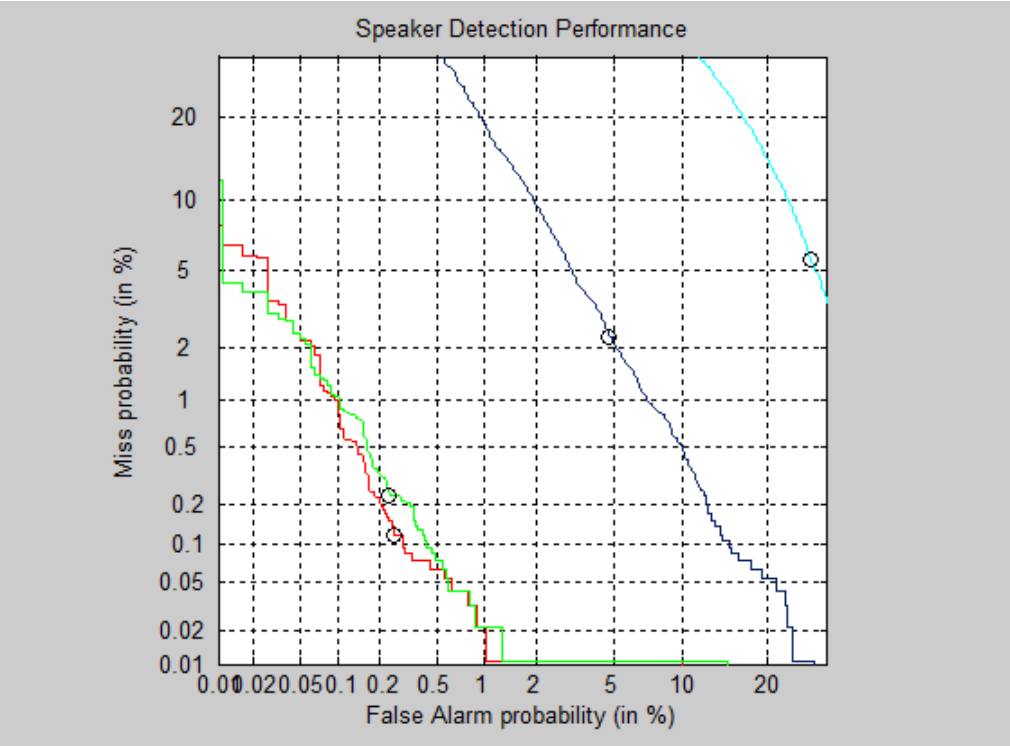

Figura 15: Curva DET del experimento de detección del sistema S2 usando los 4 juegos de modelos RPS: M1 en verde, M2 en rojo, M3 en azul y M4 en gris.

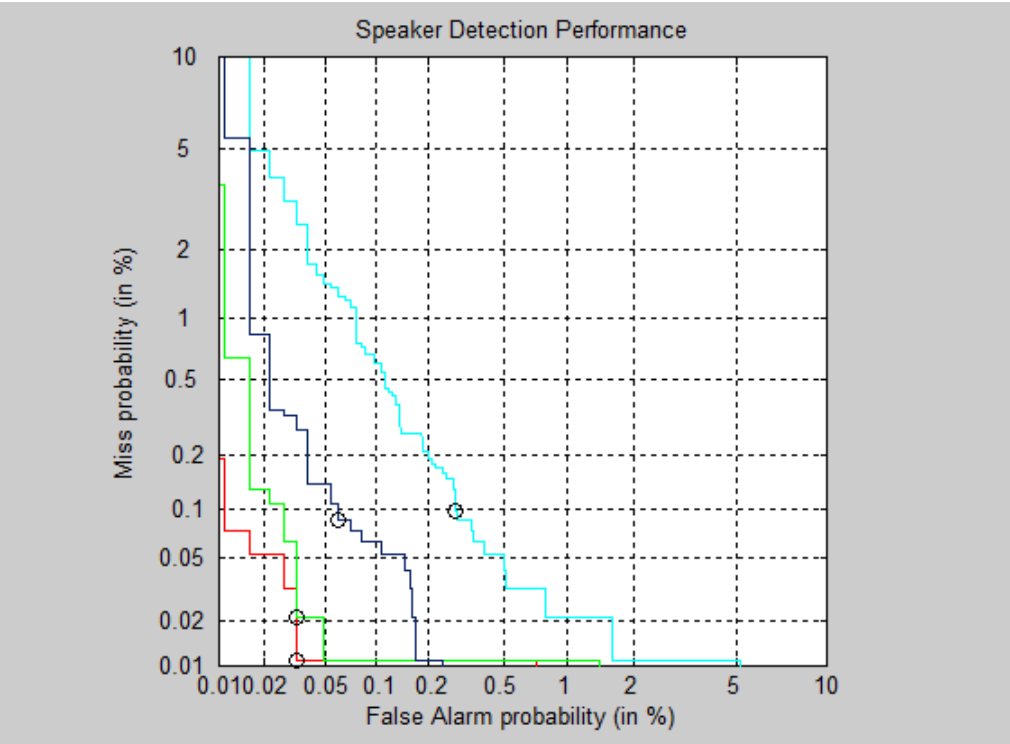

Figura 16: Curva DET del experimento de detección del sistema S3 usando los 4 juegos de modelos RPS: M1 en verde, M2 en rojo, M3 en azul y M4 en gris.

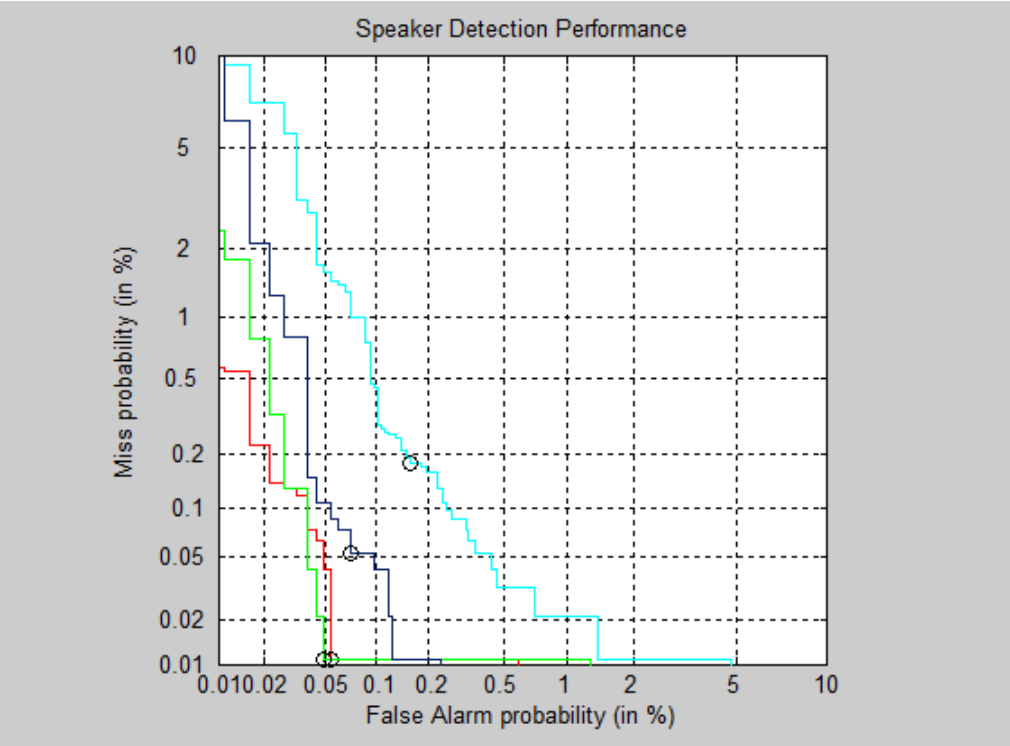

Figura 17: Curva DET del experimento de detección del sistema S4 usando los 4 juegos de modelos RPS: M1 en verde, M2 en rojo, M3 en azul y M4 en gris.

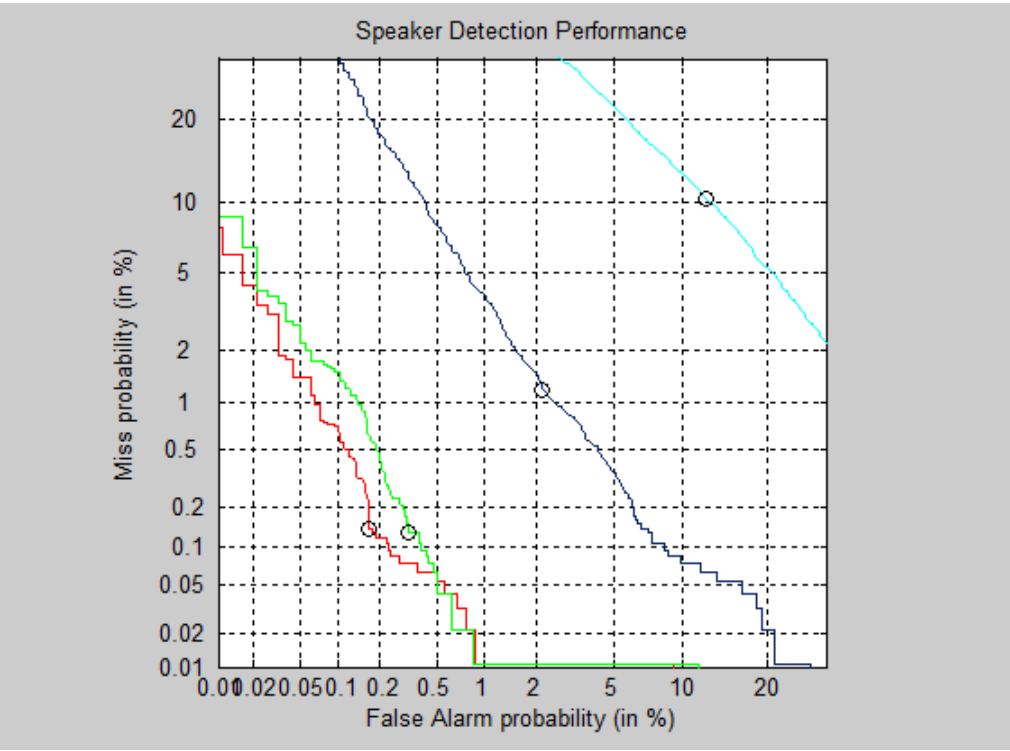

Figura 18: Curva DET del experimento de detección del sistema S5 usando los 4 juegos de modelos RPS: M1 en verde, M2 en rojo, M3 en azul y M4 en gris.

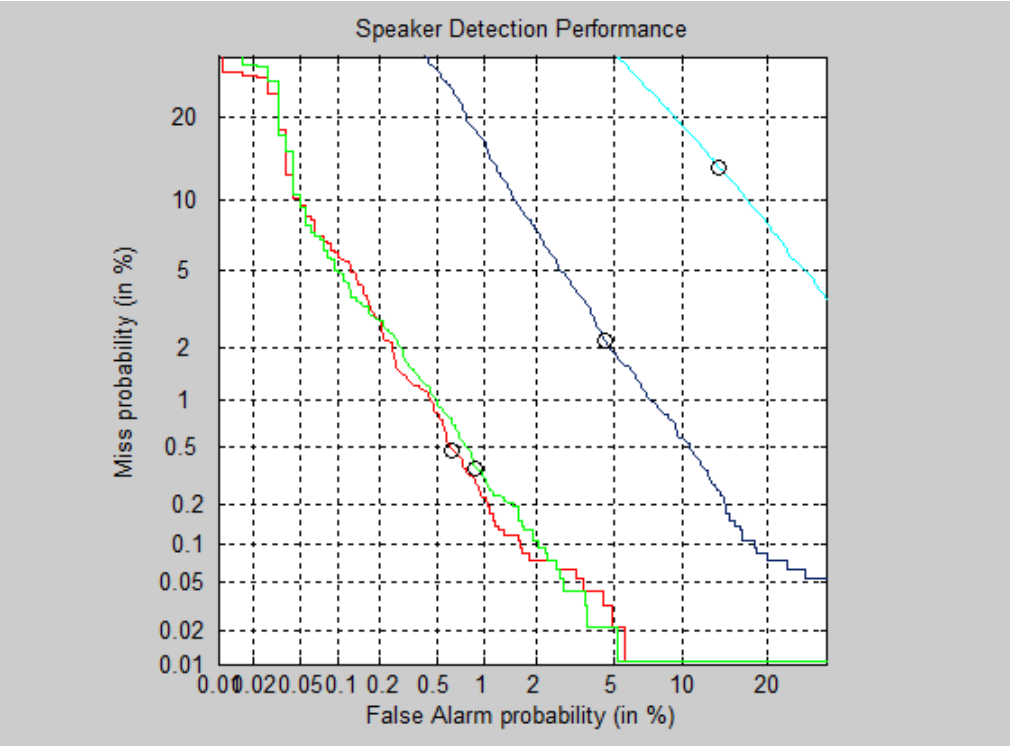

Figura 19: Curva DET del experimento de detección del sistema S6 usando los 4 juegos de modelos RPS: M1 en verde, M2 en rojo, M3 en azul y M4 en gris.

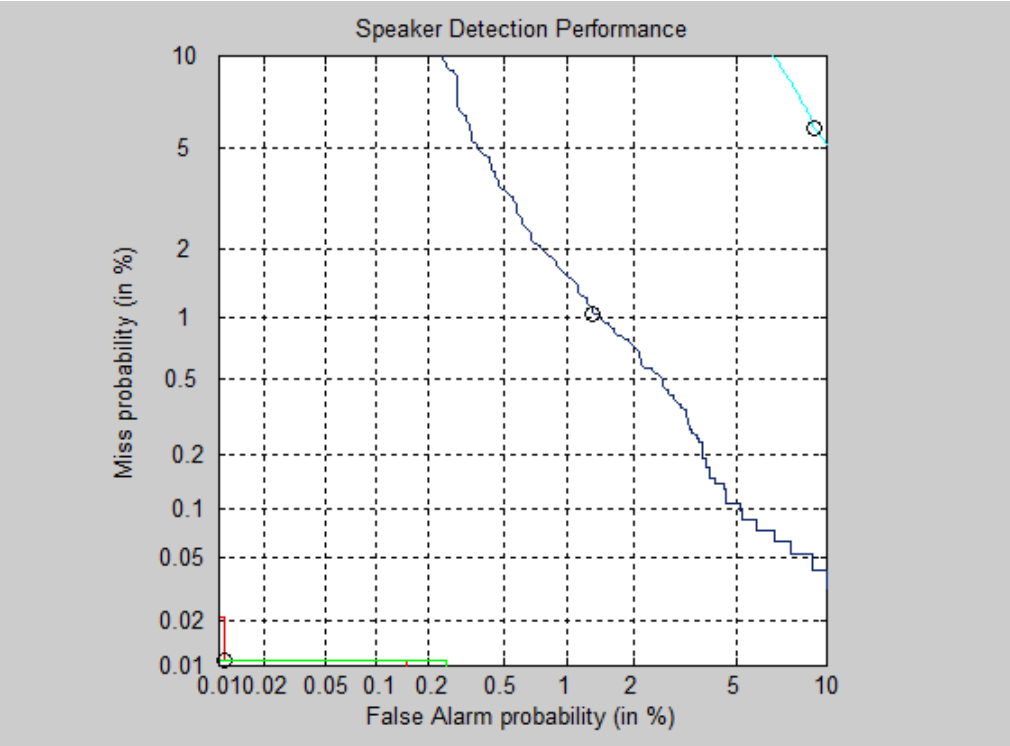

Figura 20: Curva DET del experimento de detección del sistema S7 usando los 4 juegos de modelos RPS: M1 en verde, M2 en rojo, M3 en azul y M4 en gris.

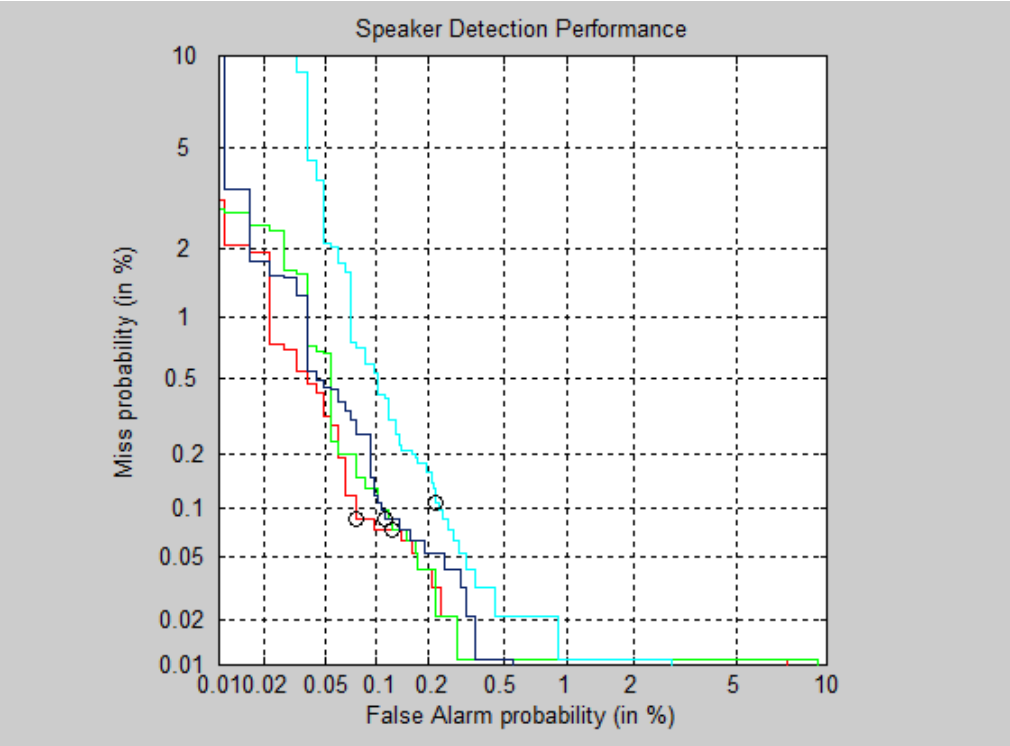

Figura 21: Curva DET del experimento de detección del sistema S8 usando los 4 juegos de modelos RPS: M1 en verde, M2 en rojo, M3 en azul y M4 en gris.

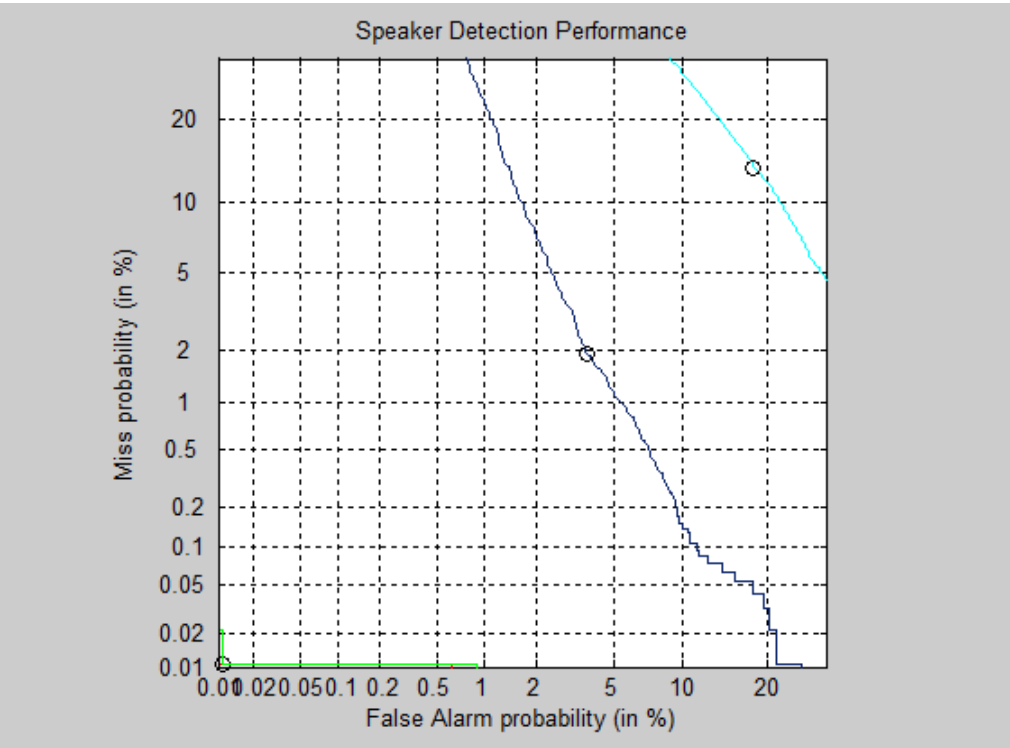

Figura 22: Curva DET del experimento de detección del sistema S9 usando los 4 juegos de modelos RPS: M1 en verde, M2 en rojo (tapada por M1), M3 en azul y M4 en gris.

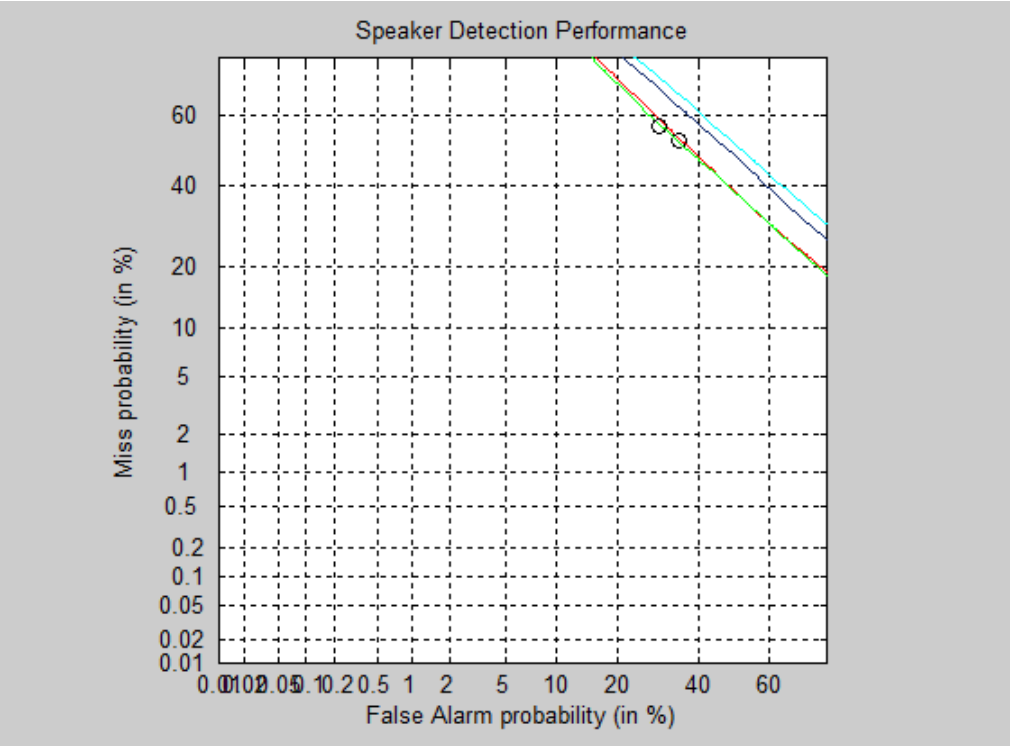

Figura 23: Curva DET del experimento de detección del sistema S10 usando los 4 juegos de modelos RPS: M1 en verde, M2 en rojo, M3 en azul y M4 en gris.

9.2.Estrategias elaboradas usando ataques específicos y copy-synthesis sobre la base de datos ASVSpoof2015

9.2.1. Evaluación usando la base de datos ASVSpoof

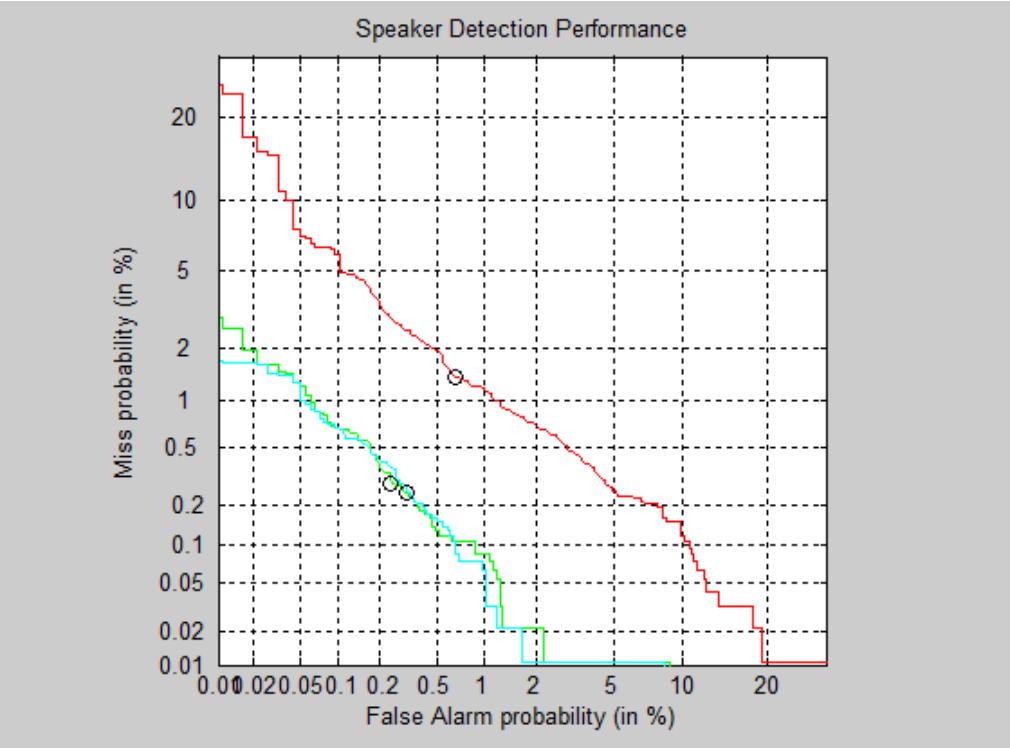

Figura 24: Curva DET del experimento de detección del sistema S1 usando modelos RPS: estrategia M1 en verde, M2 en rojo y M3 en azul.

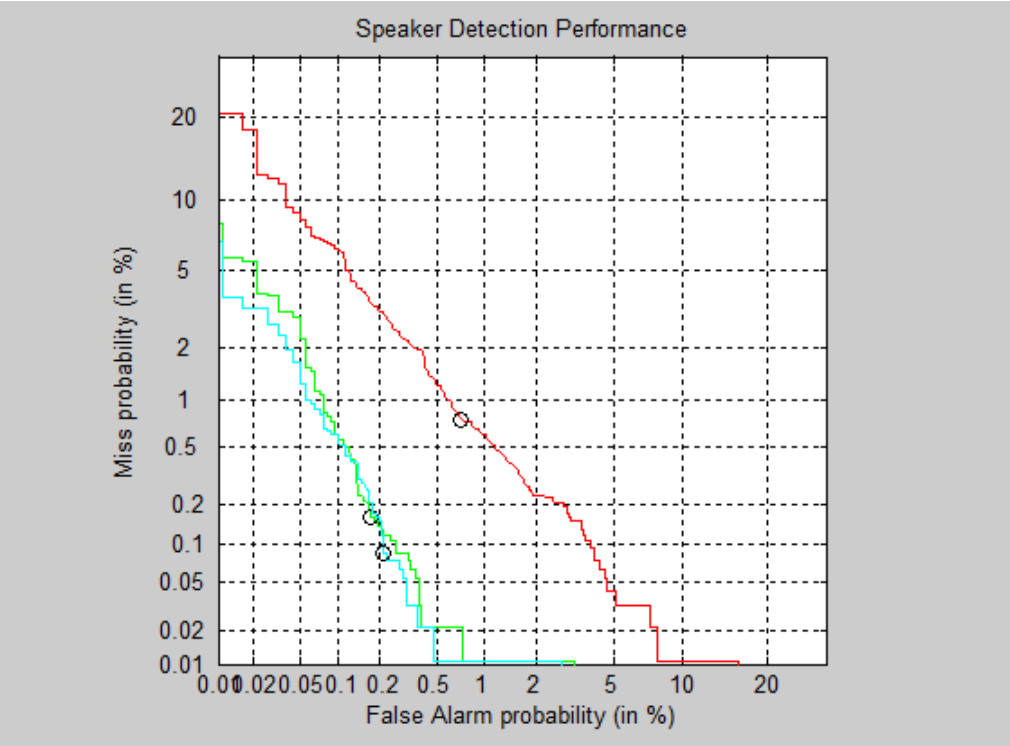

Figura 25: Curva DET del experimento de detección del sistema S2 usando modelos RPS: estrategia M1 en verde, M2 en rojo y M3 en azul.

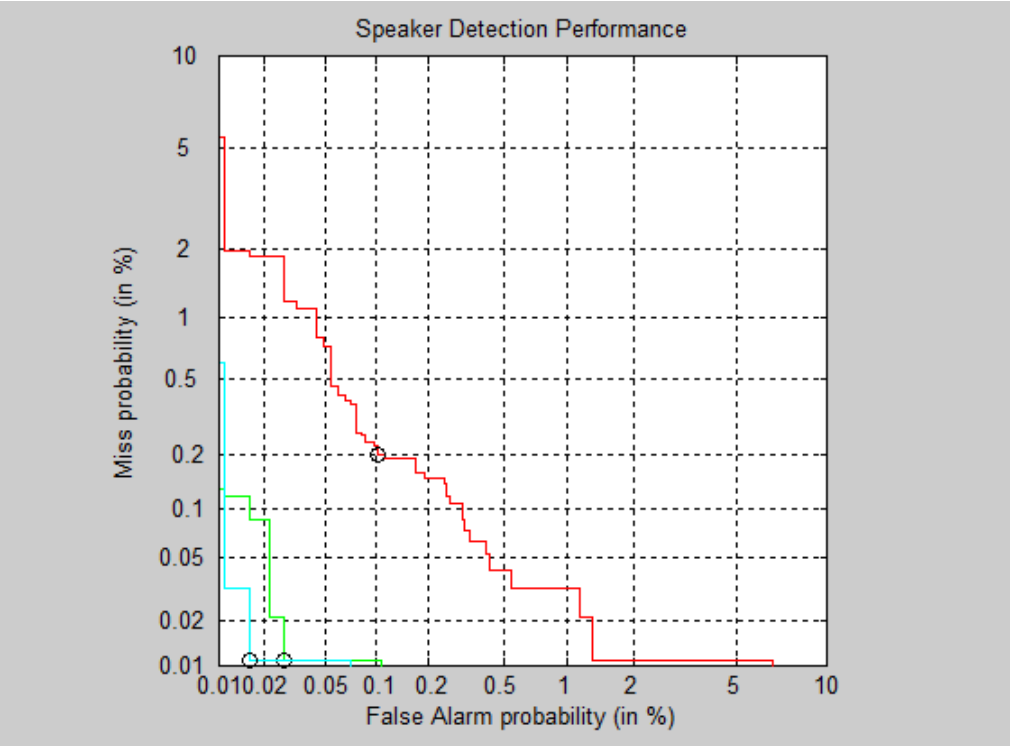

Figura 26: Curva DET del experimento de detección del sistema S3 usando modelos RPS: estrategia M1 en verde, M2 en rojo y M3 en azul.

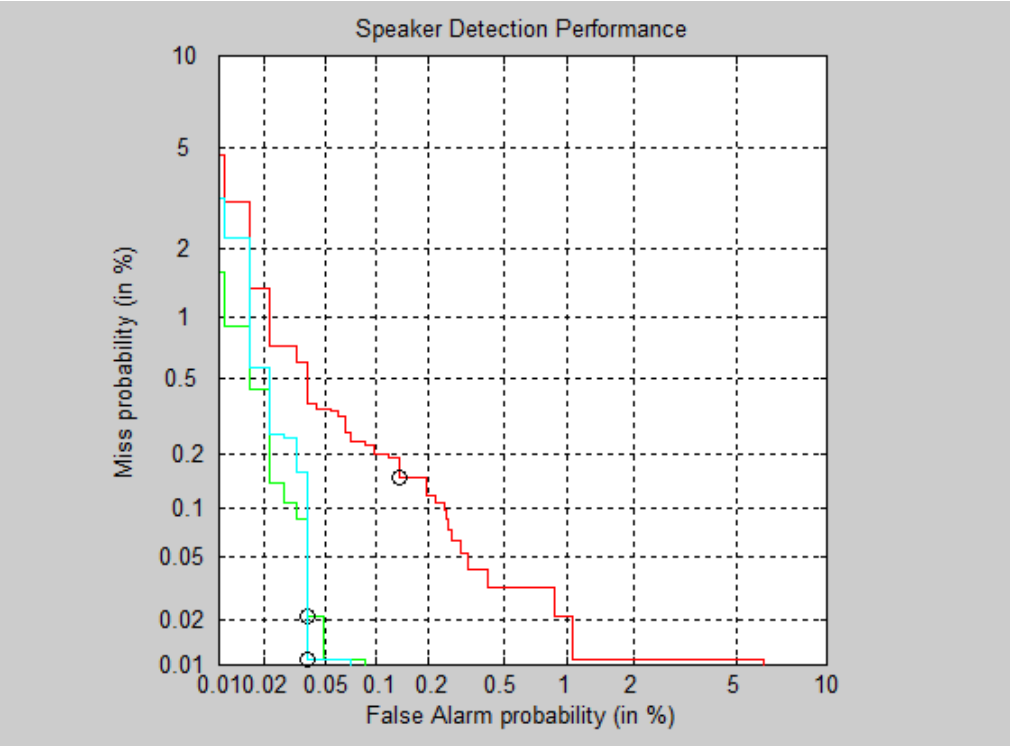

Figura 27: Curva DET del experimento de detección del sistema S4 usando modelos RPS: estrategia M1 en verde, M2 en rojo y M3 en azul.

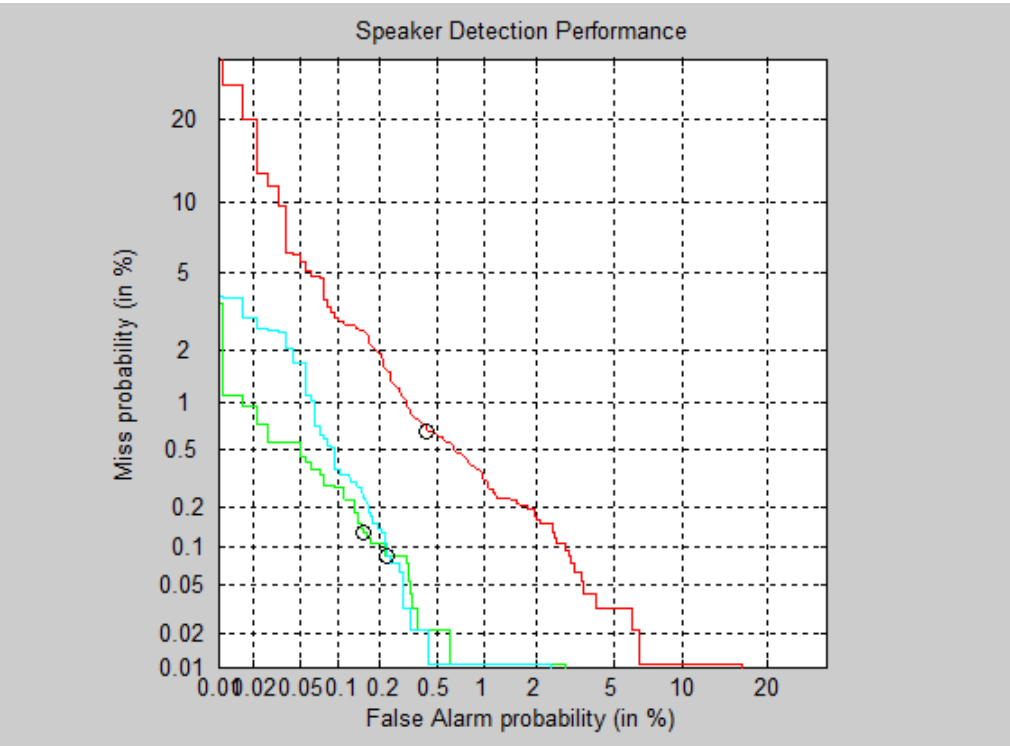

Figura 28: Curva DET del experimento de detección del sistema S5 usando modelos RPS: estrategia M1 en verde, M2 en rojo y M3 en azul.

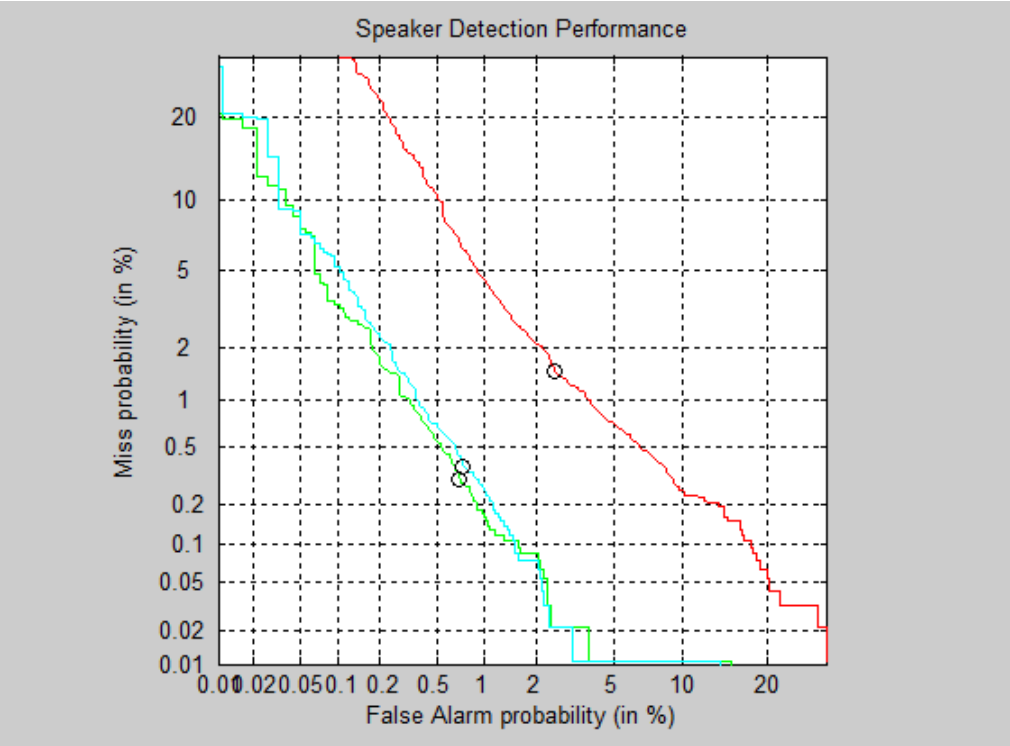

Figura 29: Curva DET del experimento de detección del sistema S6 usando modelos RPS: estrategia M1 en verde, M2 en rojo y M3 en azul.

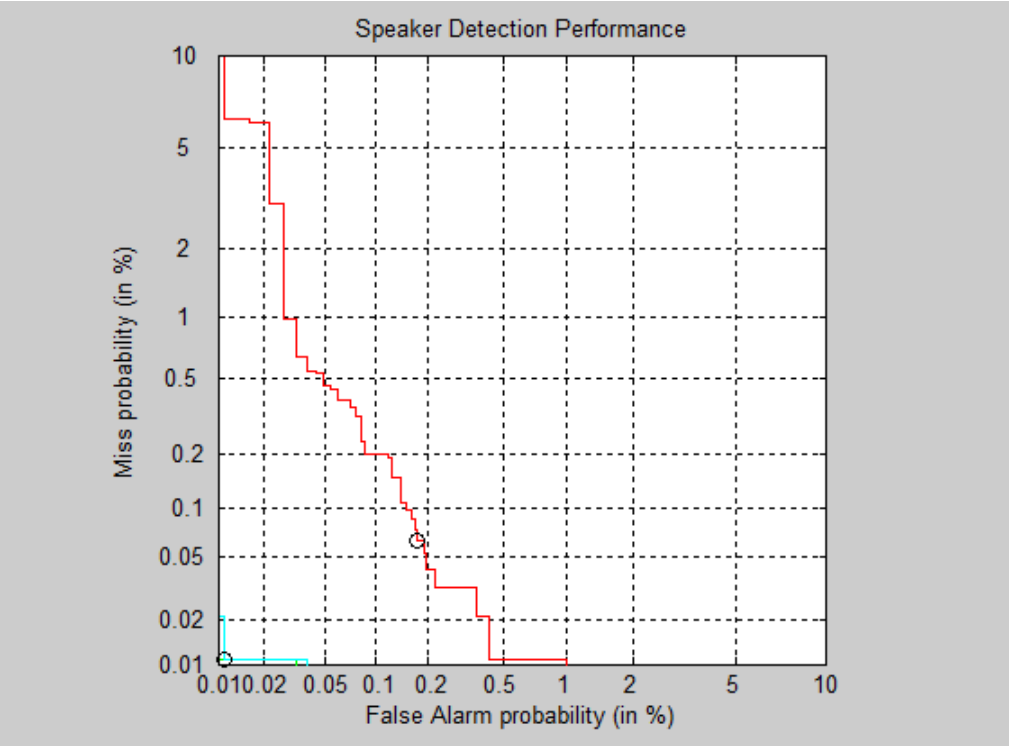

Figura 30: Curva DET del experimento de detección del sistema S7 usando modelos RPS: estrategia M1 en verde, M2 en rojo y M3 en azul.

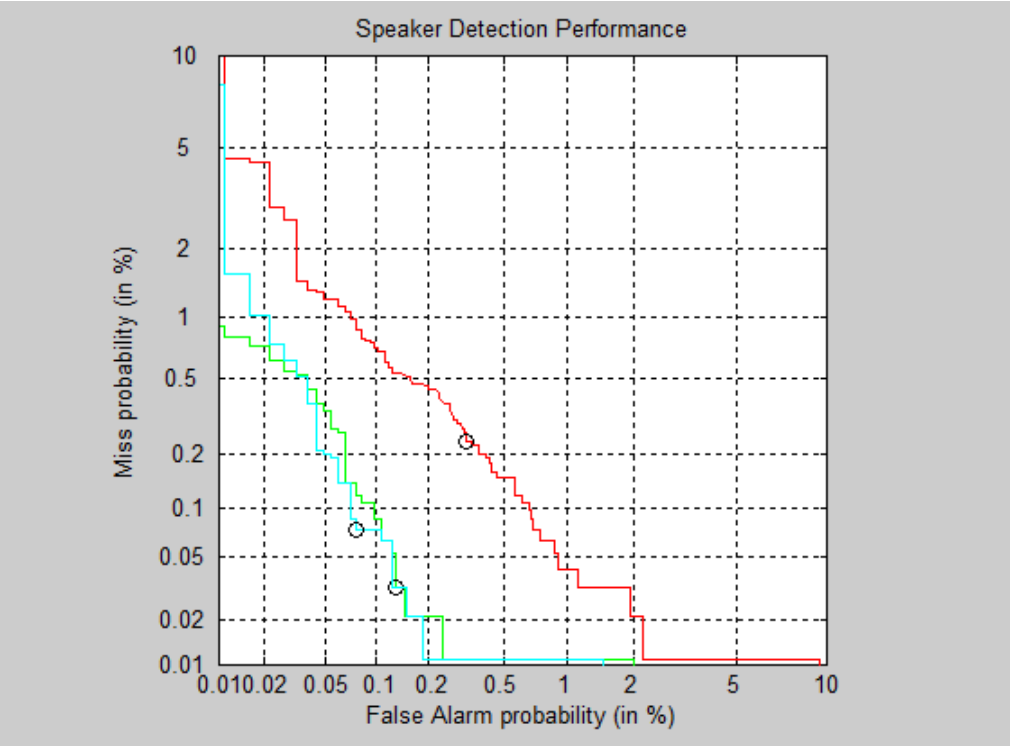

Figura 31: Curva DET del experimento de detección del sistema S8 usando modelos RPS: estrategia M1 en verde, M2 en rojo y M3 en azul.

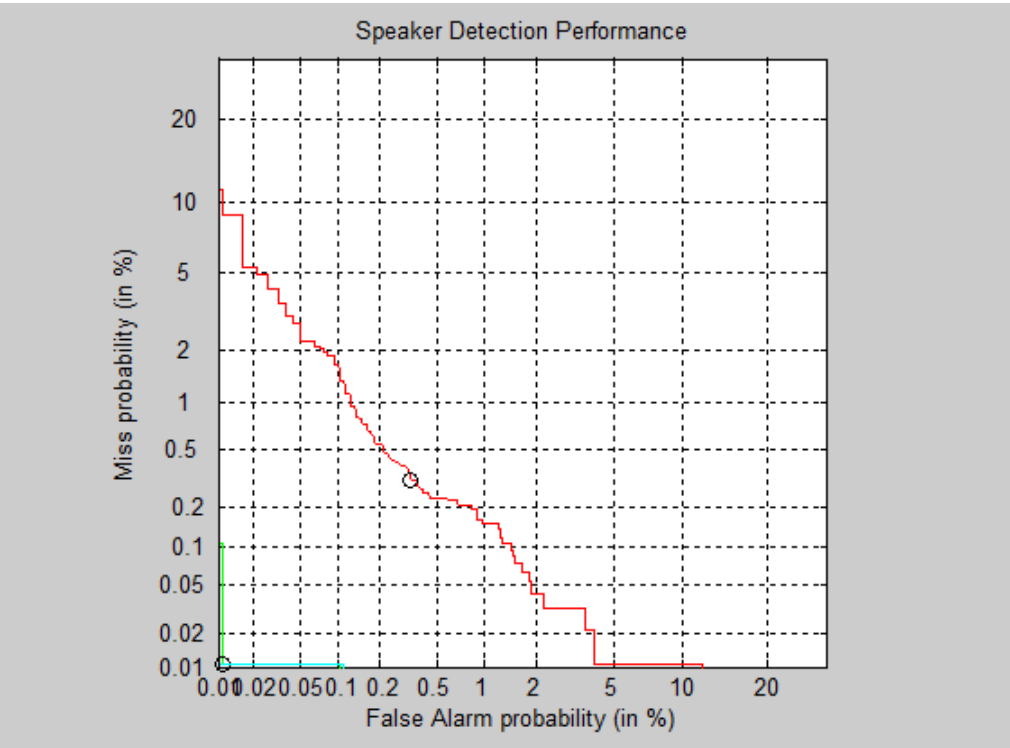

Figura 32: Curva DET del experimento de detección del sistema S9 usando modelos RPS: estrategia M1 en verde, M2 en rojo y M3 en azul.

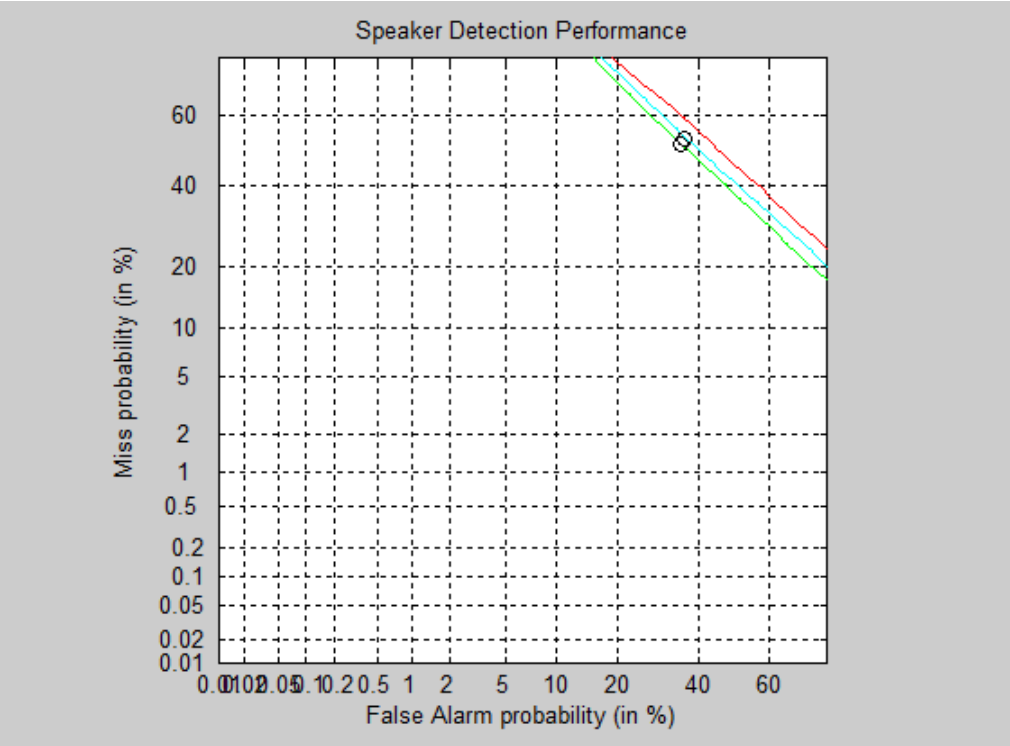

Figura 33: Curva DET del experimento de detección del sistema S10 usando modelos RPS: estrategia M1 en verde, M2 en rojo y M3 en azul.

9.2.2. Evaluación usando la base de datos de Blizzard 2012

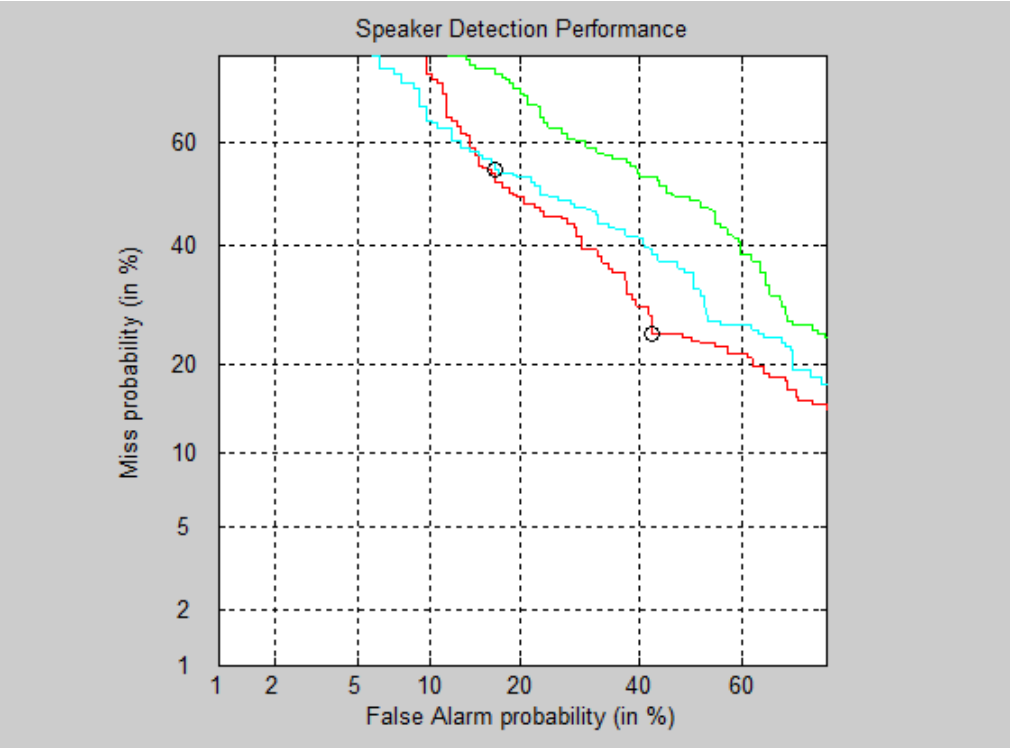

Figura 34: Curva DET del experimento de detección del sistema B de Blizzard 2012 usando modelos RPS: estrategia M1 en verde, M2 en rojo y M3 en azul.

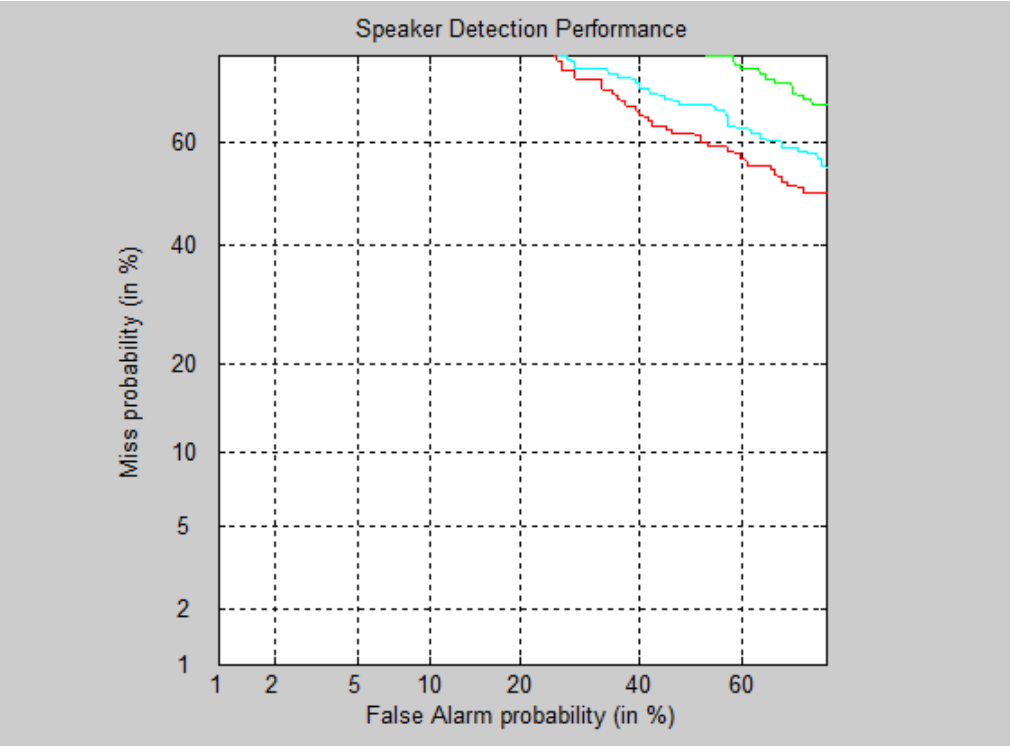

Figura 35: Curva DET del experimento de detección del sistema C de Blizzard 2012 usando modelos RPS: estrategia M1 en verde, M2 en rojo y M3 en azul.

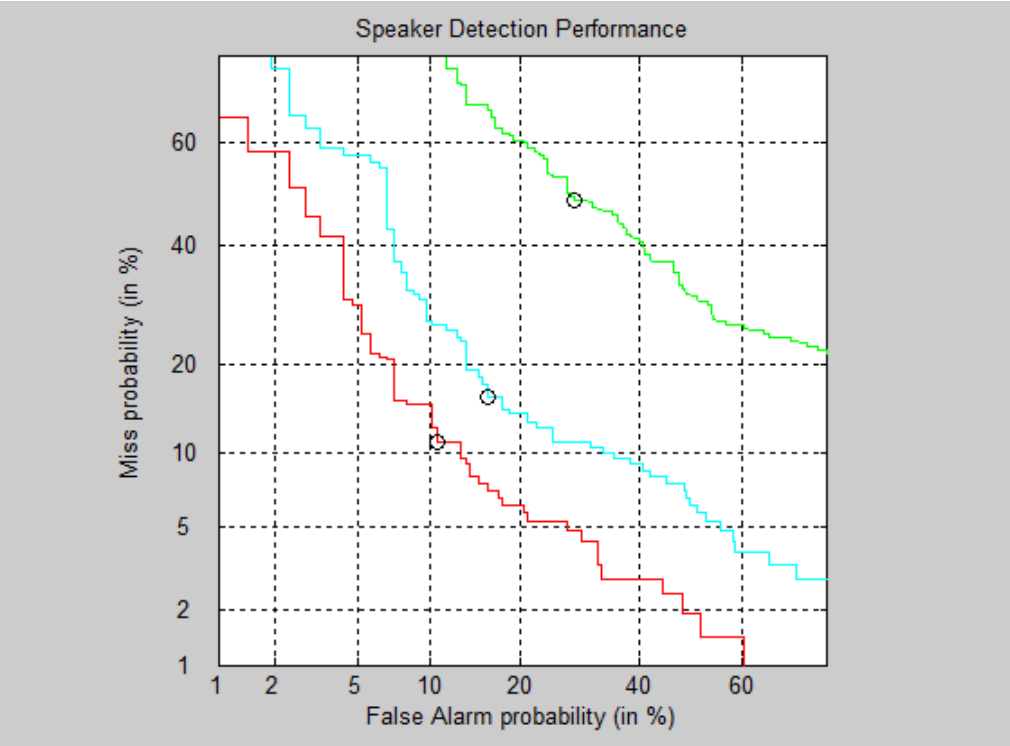

Figura 36: Curva DET del experimento de detección del sistema D de Blizzard 2012 usando modelos RPS: estrategia M1 en verde, M2 en rojo y M3 en azul.

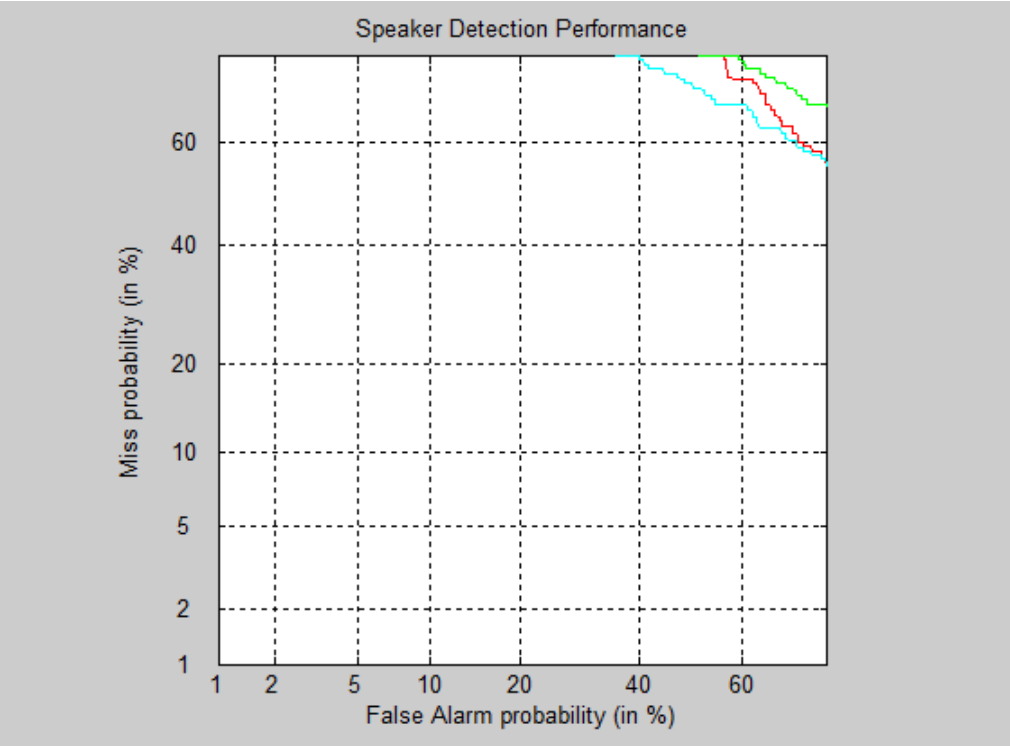

Figura 37: Curva DET del experimento de detección del sistema F de Blizzard 2012 usando modelos RPS: estrategia M1 en verde, M2 en rojo y M3 en azul.

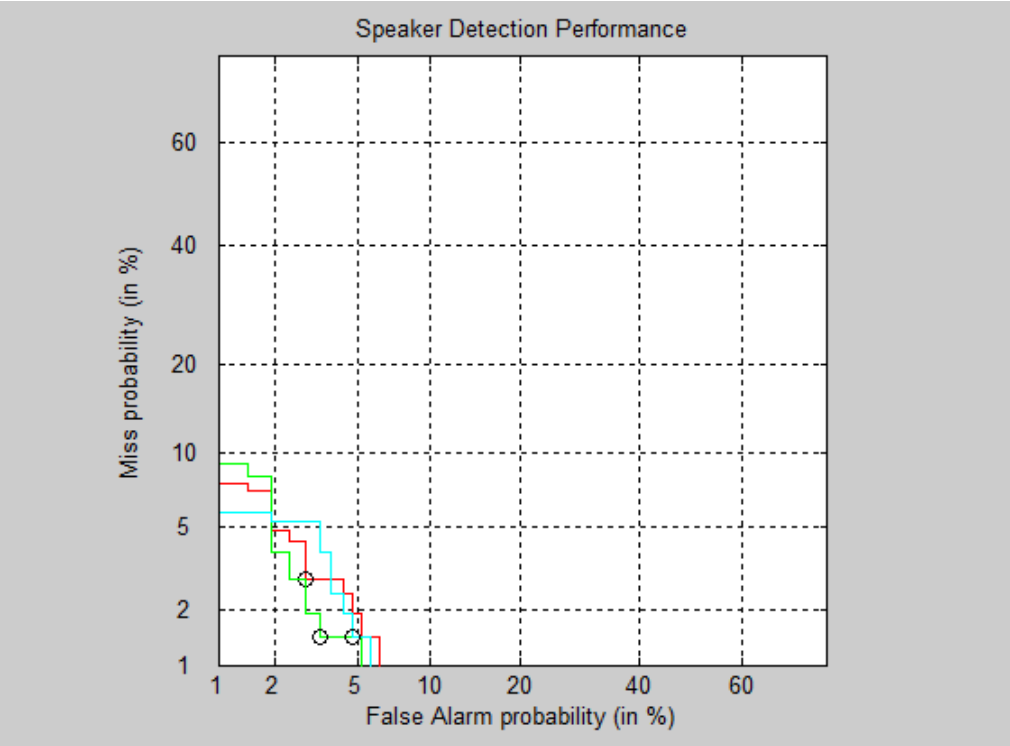

Figura 38: Curva DET del experimento de detección del sistema G de Blizzard 2012 usando modelos RPS: estrategia M1 en verde, M2 en rojo y M3 en azul.

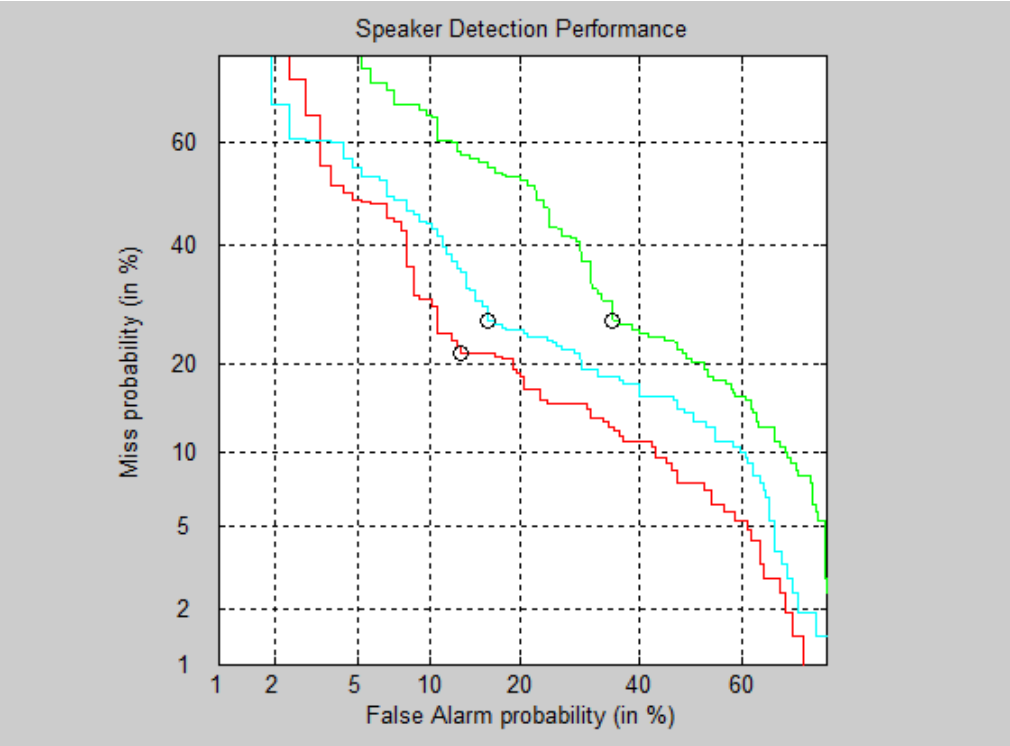

Figura 39: Curva DET del experimento de detección del sistema I de Blizzard 2012 usando modelos RPS: estrategia M1 en verde, M2 en rojo y M3 en azul.

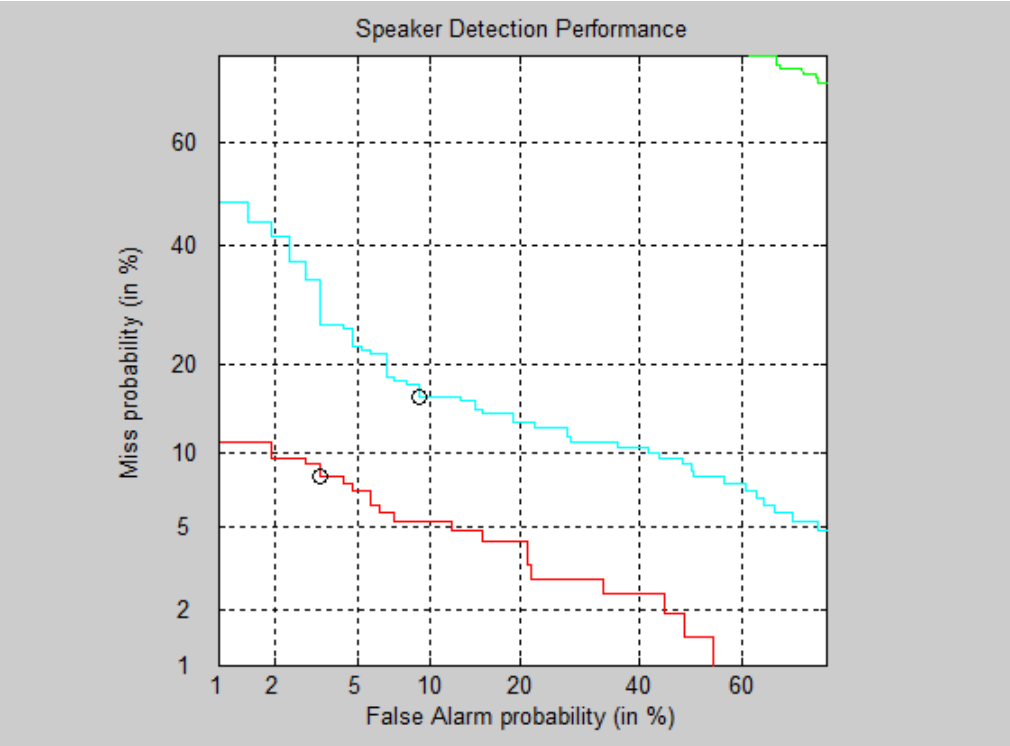

Figura 40: Curva DET del experimento de detección del sistema J de Blizzard 2012 usando modelos RPS: estrategia M1 en verde, M2 en rojo y M3 en azul.